Universitat Autònoma de Barcelona

The Web as a Corpus

A Multilingual Multipurpose Corpus

Xavier Gil Bouzou

Treball de Fi de Màster

Coordinador: Gökhan Doğru

Màster de Tradumàtica: Tecnologies de la Traducció

Facultat de Traducció i d'Interpretació

Universitat Autònoma de Barcelona, 2018/19

'The web, teeming as it is with language data, of all manner of varieties and languages, in vast quantity and freely available, is a fabulous linguists' playground.'

-Kilgarriff, A. (2001), 'Web as corpus'

Summary

The complete academic body about corpora is too extensive and wide to present it in just a few pages, therefore this work will try a different approach instead. We are in the golden age of the Web, as it is more accessible and straightforward than ever, so what would happen if we were to combine these characteristics of the Web with the traditional corpora? The result would be the 'Web as Corpus' approach. This work attempts to prove how this approach is the future of the corpora, and how these new corpora from the Web differ from the traditional corpora. To further prove the ease of use and accessibility of this approach, this work will also provide a brief and detailed explanation of how to use and create a corpus using three tools that anyone can find on the Web.

Keywords: Web as Corpus, Corpus, Web, Corpus Linguistics, iWeb, Sketch Engine, BootCaT

Resumen

La literatura académica que trata sobre los corpus es demasiado extensa y amplia como para presentarla en unas páginas, por lo tanto en lugar de ello, este trabajo opta por enfocar el tema de los corpus de una manera distinta. Estamos en la edad dorada del Internet y la web, ya que es más accesible que nunca, ¿y qué pasaría si la web y los corpus se juntaran? El resultado sería lo que se conoce como 'la web como corpus'. Este trabajo va a intentar presentar hasta qué punto este enfoque constituye la nueva dirección hacia la que los corpus van a ir en el futuro. Además, para enfatizar la utilidad y accesibilidad de los corpus, en la parte práctica de este trabajo se van a presentar tres herramientas diferentes para poder crear un corpus usando únicamente la web.

Palabras clave: la web como corpus, Corpus, Web, Lingüística de corpus, iWeb, Sketch Engine, BootCaT

TABLE OF CONTENTS

Acknowledgements

I would like to express my sincere gratitude to my supervisor Gökhan Doğru for the constructive support and insightful feedback during this challenging undertaking. Also, I would like to express my sincere gratitude to you, the reader, for giving this work a chance. Lastly, I would also like to thank my family and Moka for all the support provided.

List of Figures

List of Tables

1 Introduction

The topic of corpora is often overlooked by students, it is not a flashy topic nor an easy one to work with. But what most people do not consider is that corpora are the base of almost every work related to translation. And precisely because of these reasons I decided I wanted to work with corpora, I wanted to portray corpora as accessible and easy to use tools, not as tedious lists hard to build and work with. The problem with corpora is that they are deceivingly complex.

Before I started this study, my knowledge of the topic of corpora was null, I was not familiar with it at all, so I used this project to learn about the topic and its tradition a bit, while trying of offer an insightful approach that may bring corpora closer to students and translation professionals that may not see them as an useful tool.

With this work, I would like to bring into the spotlight the topic of corpora, as it is a topic often overshadowed by other topics (which ironically ultimately depend on corpus), as it presents a lot of 'new' interesting developments that may streamline the tedious task of creating a corpus. This work attempts to portray a new approach to corpora, a combination of the 'old' and the 'new' (the traditional corpora and the Web). In more practical terms, this study attempts to

- How the 'Web as Corpus' approach changes the notion of corpora
- Present the advantages of the 'Web as Corpus' approach
- Provide practical tools to create corpus using the Web

The structure of this work will be pretty straightforward and intuitive. This study will begin with a theoretical introduction of the topic of corpora, different definitions and approaches will be presented, to offer the user a brief introduction to the history of corpus, and a glimpse of how hard it is to pinpoint and define what a corpus really is. With all the definitions provided, the reader will be able to draw his own conclusions on the topic, or not, as there may not be a clear answer to that question. In this introductory section, not only the definition of the corpus will be provided, it will also be put in context with corpus linguistics, the academic field mostly dealing

with it (although not the only one, as it will be seen in later sections of the work…). Once the bases have been stablished, the approach of the 'Web as Corpus' will be presented and put to the test, as the advantages and limitations of this approach will be introduced shortly after. In the next section, the types of corpora will be developed, with special emphasis on the unique types of corpora that can be found exclusively in the Web. After having presented what a corpus is, and how it can interact with the Web, its new environment, it is time to start focusing on the process of creating a corpus only using the Web. This study will present three of the most renowned free tools to create corpora using the Web: *Sketch Engine*[1], *BootCaT*[2] and *iWeb*[3]. Each one of them will be presented in its own section and a step-by-step explanation on how to create a corpus using each one of them will be provided on greater detail.

2 Theoretical framework

2.1 What is a Corpus?

Before jumping into the "Web as Corpus" approach, which is the main focus of this work, it should be appropriate to provide a brief introduction to what a corpus is and how scholars have defined it, in order to establish a solid starting point.

The nature of the theoretical framework of corpora is very expansive and wide, and diverse approaches have been proposed throughout its academic tradition. It is an always-evolving topic, and one of the reasons that academics cannot agree on a universal definition is because corpora has always depended on the data that build them, which is constantly changing. And as this data and the way it is captured evolve, the corpora evolve alongside it. When it comes to a generally agreed approach on corpora, a corpus is a something that is used *"(…) to make sense of phenomena in big texts or big collections of smaller texts".* (McCarthy & O'Keeffe, 2010:3) This definition is very general, and it could even be argued that using it to define corpora would not be very representative. One could even argue

[1] https://www.sketchengine.eu/
[2] https://bootcat.dipintra.it/
[3] https://www.english-corpora.org/iweb/

that this definition could be defining various things at once. However, it cannot be denied that it encapsulates the essence of the concept nonetheless, so it is not completely wrong.

In a similar fashion, McEnery and Wilson define the corpus as *"[…] a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration."* (McEnery & Wilson 2001:32). This definition encapsulates the essence of what a corpus really is, there is no room for misrepresentation, unlike in the previous definition. Every corpus tends to have more specific connotations, they are not just compiled texts. These specific connotations that are mentioned in this particular definition are *"sampling and representativeness, finite size, machine-readable form, [and] a standard reference."*

Another definition that has been acknowledged by the academia is that of Sinclair's (1991: 171): a corpus is "*a collection of naturally-occurring languages chosen to characterize a state or variety of a language".*

The definitions provided by most authors tend to be loose in nature, and that is for a reason. As Kilgarrif and Grafenstette explain*:" […] the definition of the corpus should be broad*." (Kilgarriff & Grafenstette, 2003:334), and that is because on many occasions, the texts that linguists use as a corpus do not share many of the connotations mentioned earlier. Therefore, in order to define what a corpus really is, while avoiding some swampy semantical questions such as "*Is corpus x good for task y*?" or "*Is x a corpus at all*?", Kilgarriff & Grafenstette (2003) go one step further and suggest a modified variation of their definition: "*A corpus is a collection of texts when considered as an object of language or literary study*." What they did with this definition is to provide another perspective about corpora: "(*…) the notion of 'corpus-hood' was implicitly shifted to the intention of the researcher rather than being seen as intrinsic to the text collection itself."* (Gatto, 2013: 40)

2.1.1 Corpus Linguistics

The concept of corpus linguistics should ring a bell by now, as it is the academic field that originally dealt with corpora. To ensure a smooth and progressive understanding of the topic, now the major principles of corpus linguistics will beprovided, to further develop the topic. As it has been hinted before, *"(…) it is a mistake to assume that the analysis of corpora has nothing to offer to generative theory in particular or to theorizing about language in general*." (Meyer, 2002:4) The most popular approach to corpus linguistics is the Firthian and neo-Firthian approach to linguistics, which focuses on language not only in its linguistic context, but also in its social context. Corpus linguistics objective is to describe all the empirical aspects of language use, and to do so, Aarts (1993:3) formulated a number of requirements for a descriptive model of language, which are the following:

> *-The model should allow the combination of a quantitative and a qualitative description of the data.*
> *- The model must establish a relation between phenomena that are external to the language system and system-internal phenomena.*
> *- The model should allow the description of the full range of varieties, from spontaneous, non-edited language use (usually spoken), to nonspontaneous edited language use (usually written or printed).*
> *- The model should allow an integrated description of syntactic, lexical and discourse features*. (Aarts in Olohan, 2013:424)

Corpus linguistics attempt to seek linguistic answers by means of qualitative and quantitative analysis. The use of the Web in corpus linguistics adds new possibilities to the research, it enables *"(…) making generalizations about language use, stressing that interest is typically not just in what occurs but in what is probable and what is likely to occur."* (Kennedy, 1998:1).

2.2 The Web... as a Corpus?

As an object of study, a corpus generally "*is usually of a size which would not allow manual investigation but requires the use of specific tools to perform a quantitative and qualitative analysis of the data*" (Gatto, 2013:7) and that is the main reason technology has always been linked to corpus linguistics. Being able to perform significant empirical researches about word frequencies, patterns and comparative statistics would be practically impossible without computers. It was in the early 1960s when computer-based linguistic studies took off, with the one-million-word Brown corpus. In the early 1980s however, Sinclair went one step ahead and built an eight-million-word corpus. As technological limitations were surpassed as time went on, the limitations of corpus linguistics were also overcome simultaneously. In the early 1990s the British National Corpus (BNC) was developed and built, and it compiled 100 million words, allowing the measurement of data that was unthinkable in earlier times with earlier corpora. And even this 100-million-word corpus found its own set of constraints, such as limitations to what studies could empirically confirm, "*For some areas in corpus-linguistics, even the new mega-size corpora of the BNC-type are still not large enough"* (Hundt, Nesselhauf & Biewer, 2007:1)

Barely 10 years after the introduction of the BNC, the largest corpus until then, in 1999 Lawrence & Giles estimated that there were 800 million indexable Web pages available then. Even today it is even hard to picture 800 million pages, each one of them with its own contents and texts. Nowadays, it is estimated that there are at least around 5.13 billion pages on the Web. There are occasions when quantity may not be over quality, though. Or as Kilgarriff claimed in his paper presenting the 'Web as Corpus' approach for the first time *"Compared to the Web, the BNC is an English country garden*." (Kilgarriff, 2001:342).

The broad heterogeneity of the Web, added to the immense number of materials may be a double-edged sword, as the previous quote suggests. Kilgarriff presented three counterarguments for the 'Web as Corpus' approach, which ironically, he had introduced himself. The first problem that Kilgarriff identified was that not all documents contain text, there are pages build only with images or videos. The second counterargument he proposed was that the Web is constantly changing

so it can be hard to establish its limits. The third and last counterargument he proposed was that the Web may contain duplicates and there may be pages that can contain more than one language, which may alter the final results of a study (considering the study is dealing with a monolingual corpus).

It is true that the corpus academia owes a lot to the BNC, it was a state-of-the-art corpus that opened numerous directions for researcher, but now the world has a tool that was not present then, and that is the Web. The virtual construct known as the Web has given its users access to a myriad of different webpages and contents about a wide range of different fields, just a click away, and mostly for free. As Kilgarriff claimed: "*While the BNC and other fixed corpora remain of huge value, it is the Web that presents the most provocative questions about the nature of language" (Kilgarrif, 2001: 344).* One of the most important traits of the Web is its '*connectedness*' (Schäfer & Bildhauer, 2013:8), and it is not hard to find evidence vouching for this point, its own name says it all: *World Wide* Web.

Historically, it does not come as a surprise that the Anglophone content has completely overshadowed other languages in terms of contents found on the Web. It is important to take into consideration that the amount of content is not subject to the number of users, as it will be seen in Table 2.

See Table 1 for the ranking of the most used languages in the Web.

Top Ten Languages Used in the Web - December 31, 2017
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	World Population for this Language (2018 Estimate)	Internet Users by Language	Internet Penetration (% Population)	Internet Users Growth (2000 - 2018)	Internet Users % of World (Participation)
English	1,462,008,909	1,055,272,930	72.2 %	649.7 %	25.4 %
Chinese	1,452,593,223	804,634,814	55.4 %	2,390.9 %	19.3 %
Spanish	515,759,912	337,892,295	65.5 %	1,758.5 %	8.1 %
Arabic	435,636,462	219,041,264	50.3 %	8,616.0 %	5.3 %
Portuguese	286,455,543	169,157,589	59.1 %	2,132.8 %	4.1 %
Indonesian / Malaysian	299,271,514	168,755,091	56.4 %	2,845.1 %	4.1 %
French	412,394,497	134,088,952	32.5 %	1,017.6 %	3.2 %
Japanese	127,185,332	118,626,672	93.3 %	152.0 %	2.9 %
Russian	143,964,709	109,552,842	76.1 %	3,434.0 %	2.6 %
German	96,820,909	92,099,951	95.1 %	234.7 %	2.2 %
TOP 10 LANGUAGES	5,135,270,101	3,209,122,400	62.5 %	1,091.9 %	77.2 %
Rest of the Languages	2,499,488,327	950,318,284	38.0 %	935.8 %	22.8 %
WORLD TOTAL	7,634,758,428	4,159,440,684	54.5 %	1,052.2 %	100.0 %

Table 1. Top Ten Languages Used in the Web

Historically, the English language has always been the *lingua franca* of the Web, but that does not mean that it is the predominant language in it. (Over two-thirds of the pages written in the Web are in English [Grefenstette & Nioche, 2000]) As it can observed in Table 1, the growth of non-Anglophone users is considerably higher than English. It is a portrayal of the effects of globalization and the plurilingualism of the Web. According to Table 1 the most widely used languages are English and Chinese, clearly superior to all other languages. Considering the actual number of Chinese users and their growth rate, it would not be surprising that it surpasses English in the future. Another piece of data that is relevant to mention is the emergence of the Arabic language. It is an emerging language in the Web, with growth rates exponentially higher than the other languages. The changes in the language environment of the Web is discernably a representation of the changes of world's society at large.

The Web serves as a multilingual environment that stimulates the creation and inclusion of contents from all types of languages. "*The Web is an eclectic medium, and this is seen also in its multilinguistic inclusiveness. Not only does it offer a home to all linguistic styles within a language; it offers a home to all languages – once their communities have a functioning computer technology*." (Crystal 2006: 229). It is evident that economic factors play a role in the distribution of languages, as poor countries will have less chances to access the Web. In addition to the economic divergence, another inconvenience preventing the equal access to the Web is the encoding. Encoding non-Latin alphabets using a system specifically devised for Latin alphabets makes it even harder. (Crystal, 2006, in Gatto, 2013:53).

The following Table 2 represents the geographical distribution of the internet population worldwide.

WORLD INTERNET USAGE AND POPULATION STATISTICS JUNE 30, 2018 - Update						
World Regions	Population (2018 Est.)	Population % of World	Internet Users 30 June 2018	Penetration Rate (% Pop.)	Growth 2000-2018	Internet Users %
Africa	1,287,914,329	16.9 %	464,923,169	36.1 %	10,199 %	11.0 %
Asia	4,207,588,157	55.1 %	2,062,197,366	49.0 %	1,704 %	49.0 %
Europe	827,650,849	10.8 %	705,064,923	85.2 %	570 %	16.8 %
Latin America / Caribbean	652,047,996	8.5 %	438,248,446	67.2 %	2,325 %	10.4 %
Middle East	254,438,981	3.3 %	164,037,259	64.5 %	4,894 %	3.9 %
North America	363,844,662	4.8 %	345,660,847	95.0 %	219 %	8.2 %
Oceania / Australia	41,273,454	0.6 %	28,439,277	68.9 %	273 %	0.7 %
WORLD TOTAL	7,634,758,428	100.0 %	4,208,571,287	55.1 %	1,066 %	100.0 %

Table 2. World Internet Usage and Population Statistics

Every element that is part of this amalgamation of data that we identify as 'the Web' is interconnected to some other element, it is very uncommon to find an isolated element amidst all the data highways (this in turn, may be considered a detriment, depending on the particular perspective). Another of the main traits of the Web is its ever-expansive nature. It is not too far-fetched to say that the Web grows with each second, constantly getting new data in all shapes and forms, exponentially enlarging the Web by the minute. This data can present itself as a newly created websites, a mail from a student to her professor, or a simple ad. The case is that, in this moment, the size and extent of the Web cannot be doubted. Using the arguments of the connectedness and the sheer size of the Web would already be very valid to advocate for the "Web as corpus" approach. If this 'connectedness' and its constant increase are combined, the Web turns into an invaluable linguistic provider for translation students and professional translators, as now they have access to a tool offering almost endless possibilities.

The Web is, if used properly, *"(…) a fabulous linguists' playground"* as Kilgarriff and Grefenstette (2003:1) claim in the opening lines of their introductory paper *"Introduction to the Special Issue on the Web as Corpus"*.

At this point, let us go back to definition provided by Kilgariff and Grafenstette (2003) of what a corpus is "*A corpus is a collection of texts when considered as an object of language or literary study*." If we take into account this apparently simple definition and combine it with the "connectedness" trait of the Web claimed by Schäfer & Bildhauer (2013), an interesting interaction can be inferred with the combination of these two ideas. If we acknowledge that the Web is a connected "entity", this means that it could be examined as a single unit, or a single "*collection of texts*" about different fields, but collection nonetheless (part of the definition of Kilgariff and Grafenstette). So now that we acknowledge the Web as this collection of texts, we join this idea with the second part of Kilgariff & Grafenstette's definition. Let's see this association by means of a simple entailment formula:

a. *"A corpus is a collection of texts when considered as an object of language or literary study*" p
b. The Web is a collection of texts that can be an object of study. q

If both propositions are to be considered true (they hold a True value), then the entailment that can be inferred from them is that: The Web is indeed a corpus (when used properly)

As numerous academics have already done, and for the sake of this research, this study considers the Web *"(…) as an object of language study.*", therefore, it could be viable to claim that the Web is indeed a corpus. In practical terms, when the user looks up a word, or an expression in any search engine, the user is actually using the Web as if it was a corpus. This "Web as Corpus" is not a homogeneous practice, as there are different ways in which the Web may be used in corpus linguistics. This "Web as Corpus" approach, as this study has presented, is an amalgamation of theories and proposals of different authors, and so there are different ways that it has been implemented in corpus linguistics researches. One way of implementing this "Web as Corpus" approach is by directly using the Web, by means of the direct tools it provides to any user, such as search engines.. This is the most common way of its implementation, and the one Kilgarriff originally envisioned. But there was another trend among researchers that was equally valid and still applied to this approach, and it consisted in using the Web not as a gigantic corpus as such, but as a *"a source of textual data that are downloaded and post-*

processed according to one's needs" (Ferraresi, 2009: 2), or differently put, the Web as the source of the compilation of large offline corpora, rather than the corpus itself (Hundt, 2009:2). It is true that the line separating these two different perspectives is thin, even debatable.

This interpretation of considering the Web as a corpus has also met its detractors, as Sinclair, who claimed that the "*Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective*." (Sinclair, 2005, online)

So now that the basic notions of understanding the Web as a corpus have been introduced, the most common characteristics of corpora will be introduced, and we will see how these characteristics may apply in a virtual environment.

2.3 Main Traits of Corpora

Even though there are definitions describing corpora as a collection of texts, we should consider them more than that. If we were to acknowledge a corpus only as a simple collection of texts, *"virtually any collection of more than one text could […] be called a corpus."* (McEnery & Wilson, 2001), and that would be overwhelming for researchers and linguists. For the sake of limiting the scope of corpora studies, linguists have proposed (explicitly or implicitly) definitions of corpus *"[…] which entail fundamental criteria and standards*." (Gatto, 2013:8), such as the definitions presented earlier in this study. There are many standards across the corpus academia, but there is a fundamental agreement on a requisite; it is imperative that the texts of a corpus are authentic, representative and machine-readable (McEnery & Xiao, 2006). The other essential elements that a corpus should account for are balance, sampling, size and composition.

What is meant by authenticity is the body of a corpus should be constituted by genuine spoken and written texts. Data hailing from experimental conditions and artificial circumstances is not valid for a study whatsoever, as it will introduce artificial or modified data among the other natural results of the study. This quality is closely linked to representativeness. An authentic text, or a text that is not deliberately artificial, will be an adequate representation of the language. For example, "*texts from television interviews may appear to be natural but these are deliberately put*

under artificial conditions to get extremely odd responses" (Dash, 2015).

Another essential element that a corpus needs to account for is representativeness. In the corpora academic tradition, representativeness has been established as an essential characteristic of corpora (Francis, 1992; Biber et al, 1998). Francis understood corpora as "*a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis*" (Francis 1992: 17). On the same lines, Biber explained:

> *"A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research" (Biber, 1998:246)*

It is important to notice how 'representativeness' is the only common characteristic element presented in both definitions of what a corpus is. It is very common for the concept of "representation" to come up when attempting to define what a corpus is, as the ultimate goal of a corpus tends to be representing a language, or a part of it, in order to extract relevant and conclusive data. The goal of corpus linguistics is to find tendencies and patterns in "*language samples at the level of individual performance" (Gatto, 2013:12), s*o the authenticity and representativeness of the language informed in a corpus is what makes it such a valuable and useful tool. In other words, the results of the researches of corpus linguistics and its evolution rest on these two values, and that is why they are so important in a corpus.

When it comes to the dimension of a corpus, there is not an agreed estimated standard size. Each corpus has its own purpose and data, so it is very complicated to define what an acceptable size is. Evidently, a corpus made by a student in 30 minutes will not (and should not) be the same size of a corpus intended for research purposes. A corpus which is made for research purposes will attempt to provide various evidence of language patterns of items, therefore it will need to be sufficiently large and consistent to provide significant data as evidence. By contrast,

an instantaneous corpus tends to be relatively much smaller. As time went on and technology evolved, the size of corpora has increased exponentially. As an illustration, the first ever computerized corpus, the Brown Corpus had a size of 1 million words, and present corpora, such as the Collins Cobuild Bank of English, have 650 million words, and Web-corpora, which can amount to 1-billion words. These 1-billion corpora will probably get outdated as well, as the natural cycle dictates. An accepted truth is that the bigger the corpus is, the more data and information will contain (that does not imply that this data is desirable or relevant).

Based on what has been described until this point, the "Web as Corpus" approach sounds as a very valid modus operandi, with a lot and potential, even to the point that it would be foolish to ignore it with all the means and easiness vouching for its use. The truth is far from this utopic idea however, as this approach is not as flawless as corpora linguists would like.

The intention of this work is not challenging the traditional corpora, or proposing a new perspective, the purpose of this study is presenting both perspectives in an inclusive approach, just as Kilgarriff attempted. The "Web as Corpus" approach is another step in the evolution of corpus linguistics, it does not intend to be restrictive or exclusive, the Web *"(…) as a repository of huge amounts of authentic language in electronic format, freely available with little effort, has contributed to making the corpus linguistics approach so popular and accessible."* (Gatto, 2013:42) Ultimately, corpus linguistics and the "Web as Corpus" approach share a similar goal.

2.4 Advantages of using the Web as a source for a corpus

So why students, or professionals should use the Web as a corpus? What are the features that make it so distinct and useful for them? First, and most importantly is the fact that the Web is spontaneous and up-to-date. The content that thrives in the Web is spontaneous and self-generating (Gatto, 2013:70). It is complete and constantly building upon itself. Existing traditional corpora is finite, they have a beginning and an end, and inside it, the user can either find what s/he is looking for or not. If the information s/he seeks is not contained within that corpus,

the corpus is not useful for the task at hand, and that is not what a student or a professional needs. The Web however, may present enough examples of an expression or a construction, and is self-productive (blogs, wikis, forums…). If we had to put it colloquially, everything is possible on the Web. Users can find corpora of every language there is, even of languages that have not been compiled yet. On the Web, the user could even find corpora about the Klingon language (a constructed language spoken by the fictional race of Klingons in the fictional Star Trek universe). The vastness and inclusiveness for minority languages in the Web are welcoming factors (Baroni & Ueyama, 2006) that may not be present in any other style of corpora, and that is one of the things that sets this approach apart.

Another reason advocating for its use, as it may be presupposed and taken for granted in this 21st century, is its cost and convenience of use. The Web is virtually free, users have access to most of it without having to pay nothing whatsoever, and it can be accessed from anywhere, nowadays more than ever with the 5G technology. From desktop computers, to mobile phones, to wrist-watches, students and professionals can access the Web seamlessly and effortlessly. The construction of a corpus with materials found in the Web is cheap and fast, even to the point that the corpora built in the Web within a day or less may come close in terms of vocabulary and genres to traditional corpora, such as BNC. Sharoff (2006) and Ueyama (2006) discuss, in their respective studies, how a corpus built from the Web compares to a traditional corpus such as the BNC in terms of effectiveness. The conclusion is that in terms of vocabulary and genres, they are both very close, the difference is almost imperceptible. However, there is one significant dichotomy between the two styles of corpus, and that difference is that a Web-built corpus will "*tend to reflect more recent phrases of a language than traditional corpora, that are often subject to a certain lag between the time of production of the materials that end up in the corpus and the publication of the corpus itself"* (Baroni & Ueyama, 2006).

Another particularity unique to a corpus built from the Web are its sources and some of the genres that it can cover. One of the core principles of the Web, as mentioned previously, is its connectedness and the communication between its users. This has allowed the appearance of new 'genres' and styles of communication unique to this environment, and the only tool that allows for their

study is a corpus created with these materials only found in the Web. In the Web the user can easily find personal blogs, discussions, reviews, debates of any kind. All of these are genres that can be found almost exclusively in a virtual environment. Another advantage is that due to this interactive nature, the samples for these corpora may possess some characteristics of oral communication (Storrer & Beißwenger, 2008), one of the most underrepresented text in corpus linguistics, due to the complications of acquiring and registering this type of data (Lew, 2009:8). The representatives of the Web cannot be found anywhere else, it grants the users access to a wide range of written genres which are progressively growing and naturally expanding. And what's more, in view of the strong progress of the social media, linguists, and users in general are able to learn more about discourses not seen before, as natural everyday exchanges, telephone conversations and the like (Leech, 2007:144-5).

2.4.1 Limitations of using the Web as a corpus

But the use of corpora built with the help of the Web does not only imply advantages, as it also has its *drawbacks*. If this approach only had advantages, then there would be no need to compare and research new methods ("if it's not broken, don't fix it"), and this discussion would be pointless, so the evolution of corpus linguistics would eventually come to a halt.

One of the main problems of using the Web directly as a corpus is the lack of information and balance during the text gathering process. Probably, the person in charge of constructing a Web-based corpus will not have full control over all the sources that will end up building the corpus. The open and inclusive nature of the Web may be a double-edged sword. In the process of gathering and compiling data, there may be some collection of texts that are "*unplanned, unsupervised (and) unedited*" (Gatto, 2013:43). The user cannot read one by one all the texts included in an online corpus, it would be too time consuming. Also, it is almost impossible to keep track of all the text types and genres it contains; therefore, the quality of the corpus material may be dubious.

It will also probably enclose repetitions and other unpredicted elements that may affect the results. Or in other words, the corpus may contain "noise". The "noise"

of a corpus created from the Web can present itself differently. It can appear as duplicates, repetitions, all the non-linguistic materials that will eventually alter the final results, and typos. The errors and mistakes of the materials are particularly common in the materials in the Web. Kilgarrif and Grefenstette were not contrary to the inclusion of these errors, "(…) *the Web is a dirty corpus, but expected usage is much more frequent than what might be considered noise*." (Kilgarrif & Grafenstette, 2003: 342). They believed that these mistakes, instead of being ignored and avoided, they could be consciously acknowledged in the final results (not always, though), as they could be of interest for other fields such as EFL or sociolinguistics, among others. It is very important for corpus linguistics to become aware of them and profit from them.

There are some practices that produce noise that can only take place in the Web and are therefore exclusive to the corpora created from the Web. Some Web content creators, for instance, insert in their webpages a high number of repeated keywords in a way that is unnoticeable for the common user (by altering the code, or using an imperceptible font in the background. This practice was coined as Web-spamming (Gyongyi & Garcia-Molina, 2005), and its objective is boosting the position of said pages on the order of results when using a search engine. This over-repetition may not only cheat the search engine algorithm, it will also create a misrepresentation in the results of the corpus.

Also, *"(…) due to the ephemeral nature of the Web, replicability of the results is impossible*." (Hundt, 2009:3). For better or for worse the Web has turned into a commercial Eden. It is regrettably biased, as it is mostly influenced by this emerging cyber-capitalism (those who have the money control what the user sees). Nowadays we are in a near-dystopic environment where the big corporations try to hide how Google or Facebook for example, register and save secret data from the users, even the government of the United States of America does it. And how is this connected to the construction of a corpus? During the process of data-mining, when the corpus is incorporating all the data necessary for the compilation, the Web will keep track of the location of the user, and algorithms will adjust to the preferences of the user, and "bias" the search results. Hence the results of the study will be different depending on the location of access to the Web (Fletcher, 2005). Nowadays the algorithm of the search engines tends to rank the search results by link popularity,

which favors the appearance of a relevant link among the top search results, but also favors the social influence of popular commercial sites.

Another disadvantage of corpus linguistics is the restrains related to the copyright of the resources. That may not be considered a practical inconvenience *per se*, but a legal and administrative one. It is very important to bear it in mind while creating a corpus nevertheless. If a corpus created from the Web, which contains 10M documents from the Web for example, the procedure of getting all the permits and certifications from all the copyright holders may last longer than creating and publishing a traditional corpus.

It is important to note that this inconvenient is not exclusive to the "Web as Corpus" approach, it involves the whole area of corpus linguistics, due to its data-dependent nature. Kilgarriff and Grafenstette (2003) were the first ones to defend their approach, and regarding this matter, they said:

> "*Lawyers may argue that the legal issues for Web corpora are no different from those around non-Web corpora. However, first, language researchers can develop Web corpora just by saving Web pages on their own computer without any copying, thereby avoiding copyright issues, and second, a Web corpus is a very minor subspecies of the caches and indexes held by search engines and assorted other components of the infrastructure of the Web: If a Web corpus is infringing copyright, then it is merely doing on a small scale what search engines such as Google are doing on a colossal scale.*"
> (Kilgarrif & Graffenstette, 2003: 355)

So, the Web was considered less constrained in terms of copyright and legal obligations. Compared to what the established search engines are doing everyday with the information they gather, the use of Web data to build a corpus is not even significant. If the copyright holder of the content notifies an infringement, that material will have to be removed from the sample, as dictated by the *Online Copyright Infringement Liability Limitation Act*. There are some solutions to bypass the copyright problem.

The first one, and the most traditional one is requesting the copyright holders their permission to use and redistribute their texts, obviously this option will be discarded when talking about a Web corpus, as asking permission to millions in potential copyright holders for usage would be impossible (Baroni et al. 2009:18). The second approach, also quite indisputable, is using exclusively sources that are catalogued as public domain. Gatto proposes a third approach, "*to collect data regardless of copyright infringement but avoid distributing them*." (Gatto, 2013:64). If the data can only be accessed through online interfaces, the legal responsibility shifts away from the user. This is the most common approach at the present, most of the corpora created with contents in the Web are accessible through indirect sites, such as corpus.byu.edu (Davies, 2005) or Sketch Engine (Kilgarriff et al., 2004)

Another way to sidestep copyright infringements is redistributing only the Webs or links, rather than the texts. (McEnery & Hardie, 2012 in Gatto, 2013:65). If there are no texts, there is no risk of copyright infringement.

There are others, however that believe that students, teachers or researchers should not find any trouble in this regard, as it is considered that *"(…) a Web-accessible corpus for research and education derived from online documents retrieved by a search agent in ad-hoc searches will fall within legal boundaries"* (Fletcher, 2004:281). There is a very thin line that separates what it is safe to use as a resource and what not, but the legal and ethical issues go much farther than this, and a much more extensive discussion should be developed in a different study.

2.5 Types of Corpora

While it is true that every corpus has its unique purpose and identity, it is also true that corpora tend to be classified in diverse categories depending on their purpose and characteristics. Gatto (2013) presents three different benchmarks that allow to categorize corpora depending on the data they contain: "*general vs. specialized, synchronic vs. diachronic and monolingual vs. multilingual*". (Gatto, 2013:15). These categorizations tend to be mutually exclusive, that is to say, a monolingual corpus cannot be multilingual at the same time, and vice-versa.

On the one hand, a general corpus is, as its name already suggests, a corpus that contains texts from a variety of different fields, genres and registers, in order to be as representative as possible. This type of corpora is used to *"(…) produce reference materials for language learning and translation, such as grammar books or dictionaries*" (Gatto, 2013:15). As they contain widely distinct data, they could be considered as a linguistic Swiss knife of sorts. Examples of general corpora are the first computer corpus, the Brown Corpus and the British National Corpus (BNC). On the other hand, a specific corpus unlike a general corpus, will forfeit its broadness and only represent a certain field, genre, time or variety of language. Due to their specificity they are shorter than general corpora, but they are also less ambiguous and easier to use and study. Examples of specific corpora are the Corpus of Early English Medical Writing (CEEM) or the Air Traffic Control Speech Corpus (ATCOSIM).

Corpora are also distinguished between synchronic and diachronic corpora. A synchronic corpus contains data from a certain limited period of time, trying to portrait the characteristics of a language in a span of time. Some examples of synchronic corpora are the Helsinki Corpus of English Texts, or the Corpus of Contemporary American English (COCA). Diachronic corpora try to seek the evolution of a language; therefore, they include texts from all ages and periods. Inside this categorization of corpora, Sinclair (1991) proposed a special kind of diachronic corpora, labelled monitor corpora. The objective of this kind of corpora is to "monitor· the changes in the language, so new texts are continuously being added to update it, so it is constantly growing. An example of monitor corpora is the Bank of English (BoE) that is constantly updated with contemporary texts.

The last defining factor of a corpus are the languages that it includes, they can either be monolingual, or multilingual. Multilingual corpora include texts in different languages, and the most common types of multilingual corpora are parallel corpora and comparable corpora. A parallel corpus contains the same texts in two or more languages (so at least the original text and a translation), such as the open source parallel corpus (OPUS), an online corpus containing aligned corpora from diverse fields and institutions, whilst a comparable corpus contains two or more collection of texts sharing traits such as genre, topic and time span, so they can be compared. The International Corpus of English (ICE) and the International Corpus

of Learner English (ICLE)

There are numerous corpora that are accessible from the Web, with different formats, styles, genres and languages, so providing a list of all of them would take too long. Therefore, a small comprehensive list of the most renowned corpora in English is provided hereafter.

General information	
Name of the corpus	iWeb: The Intelligent Web-based Corpus
Name of the institution	BYU (Google Scholar)
URL	https://www.english-corpora.org/iWeb/
Size	14 billion words
Languages	EN
Domain of texts	General
Additional information	
The scale of the iWeb Corpus surpasses by miles other corpora that may be found in the Web, therefore it should be a good opportunity to present a practical case of how a virtual corpus may be used. Thus, it will be analyzed in more detail in the next section, in order to see what sets it apart and how future corpora, and the Web as Corpus approach have influenced it.	

General information	
Name of the corpus	British National Corpus (BNC)
Name of the institution	Oxford University press
URL	https://www.english-corpora.org/bnc/
Size	100 million words
Languages	EN
Domain of texts	General (Spoken 10m, Fiction 17m, Magazines 16m, Newspaper 11m, Academic 16m, Other 30m)
Additional information	
The BNC is 10% spoken / 90% written. The BNC has a wide range of subgenres, and covers considerably good the informal register, used in magazines or newspaper, albeit the texts are outdated.	

General information	
Name of the corpus	Corpus of Contemporary American English
Name of the institution	Brigham Young University
URL	https://www.english-corpora.org/coca/
Size	560 million words (20 million words per year)
Languages	EN
Domain of texts	General (Spoken 118m, Fiction 113m, Magazines 118m, Newspaper 114m, Academic 112m)
Additional information	
The division of texts is very similar to BNC, but the COCA is much larger and more recent, which may be a deciding factor when it comes to choosing one, as outdated data is not valid for empirical researching ongoing linguistic phenomena.	

General information	
Name of the corpus	Hansard Corpus
Name of the institution	UK Arts and Humanities Research Council
URL	https://www.hansard-corpus.org/
Size	1,6 billion words
Languages	EN
Domain of texts	Texts from the British Parliament
Additional information	
This corpus contains every speech given in the British Parliament from 1803 to 2005 (7,6 million speeches)	

General information	
Name of the corpus	NOW Corpus (News On the Web)
Name of the institution	Brigham Young University
URL	https://www.english-corpora.org/now/
Size	7.6 billion words (140-160 million words per month)
Languages	Available in multiple languages
Domain of texts	Web-based newspapers and magazines
Additional information	
The NOW corpus is the fastest-growing corpus that can be found at this moment. This corpus is the answer to those who think that corpus tend to get stale. Automated scripts run every day to add texts to the corpus. This means that all the text and all the data extracted from this corpus is representative, not only of 10 or 20 years ago, but also help to analyze rising new words or expressions. Around 9,000-10,000 new texts are added every day (Davies, 2017:2)	

2.6 Corpus-based Translation Studies

Once having presented a quite extensive and broad approach about corpus linguistics and the types of corpora that exist, the basics should be set by now. We will connect all what has been introduced about corpora and associate it to our field and topic of interest: how to translate with a corpus, or more formally put it, how corpus linguistics and translation studies may be reciprocally connected

Corpus linguistics have used corpora to approach language empirically, the goal of this approach is describing the features that are present in a language, and/or provide a qualitative study of certain elements. This empirical approach also allows to study certain features of translated text, and it even allows comparative studies on the basis of the translator's style.

The connection of translation and corpus linguistics has not always been as accepted as it is now. Traditionally, translation studies were not considered representative of the language use, thus translated texts have not always been considered as part of a corpus, "*the way in which they are used in parallel corpora indicates that translations are not seen as texts which exist and function in their own*

right in the target language system, nor as being subject to a range of constraints which differ from other text production situations" (Olohan, 2002:419).

To counter argue this disconnection, one could say that *"(…) a bilingual parallel corpus is a corpus that contains the same text samples in each of two languages, in the sense that the sample are translations of one another*" (Oakes & McEnery 2000:1), therefore entailing that since the apparition of the first bilingual corpus, corpus linguistics and translation studies have been implicitly linked, in one way or another. If in a parallel corpus there is a text in different language that means that the translated text has to be considered the "same" text. However, using translation in corpus linguistics has brought nothing but discrepancies, as no translation and translator works the same way. Different translators make different translations of the same sentence, so there may be discrepancies and criteria that may compromise the cohesion and the results in a corpus linguistics research.

Regarding the problems and difficulties of using translations in corpus for a linguistic study, Olohan offered a very thorough insight:

> *"It is well-known that linguistic choices often differ depending upon the individual translator, or there may be outright mistakes in translation. To what extent can we then make generalizations based on translated texts? And can we really be sure that the same meanings are expressed in the source and the target text? Or should we rather think in terms of degrees or types of equivalence? […] Most seriously, to what extent can we take translated texts to be representative of ordinary language use? Translated texts may differ from original texts because of source language influence […] Moreover, there may be general features which characterize translated texts."* (Olohan 2002:420).

But what is the focus of translation studies? As Gideon Toury suggested, translation studies should "*focus (the) research on anything which is assumed to be a translation"* (Toury 1995:31). It is deeply engaged in identifying or labeling something as a translation and appraise its linguistic value. However, they are not only interested purely on the linguistic value of a translation, much like in corpus

linguistics, the text *"(…) is not purely linguistic transfer conducted in a vacuum but social acts and cultural events governed by various linguistic and cultural constraints. It is a kind of cultural fact of the target language with its own distinctive features rather than the derivative of other texts."* (Hu, 2011:5)

Translation studies are expected to describe the features and roles of a translated text and translation process with reference to the political, ideological, economic and cultural contexts in which translated texts are produced. (Hu, 2011:4) And so, the ultimate goals of corpus linguistics and translation studies are not that different. They both

> *"(…) emphasize the significance of descriptive research supported by empirical evidence and the necessity of contextualization. Linguistic regularities are regarded as probabilistic norms of behavior rather than prescriptive rules. These language patterns are inextricably related to sociocultural variables insofar as they reflect and reproduce culture. It is these similarities that enable the marriage of corpus linguistics and descriptive translation studies."* (Hu, 2011: 5)

Between 1993 and 1998, scholars from Europe and North America pioneered a new way of creating and using corpora. They started to create corpora only to define the basis and significance of translations studies. These corpora would contribute to identify and establish the shared characteristics and shortcomings of translated texts and therefore, of translation studies in a broader perspective. The connection of corpus linguistics and translation studies has a reciprocal nature; corpus linguistics gets more data and more texts when using translation, therefore extending the reach of the empirical data it can provide, whilst translation studies get to develop and *"(…) enable translation scholars to uncover the nature of translated texts as a mediated communicative event."* (Baker (1993:243). Due to this apparently unlikely combination, the translations studies took a turn there, and from the marriage of corpus linguistics and translation studies grew the corpus-based translation studies. The first translational corpus was the Translational English Corpus (TEC), compiled in 1995 by Mona Baker and her team. Since this first corpus, corpus-based approach translational studies have been seen nothing but a steady rise in use and popularity.

One of the earliest advocate of this approach was Laviosa, who wrote about this first in her Ph.D dissertation in 1996 and later wrote a book about it two years later, in 1998. According to her, "*corpus-based translation studies investigate features of different kinds of translations, and the research approach it adopts is characterized by the combinative use of bottom-up and top-down methodology and by the blend of quantitative and qualitative research methodologies as well.*" (Laviosa, 1996 in Hu, 2013:7)

And this development and unification of translations studies, thanks to the corpus-based methodology, gave as a result what Baker termed "features of translation", also known as "universals of translation". But how were these "features of translation" developed? What the translation studies scholars did was use the different types of corpora that were established in corpus linguistics to identify and categorize translation-specific features that are prevalent in translations. They used parallel corpora for "*translator training and machine translation, and that made possible a shift possible a shift from prescriptive* translation research to descriptive translation research." (Hu, 2011:6). When corpus linguistics and translation studies began to be on the same page, the applications of corpora exponentially increased. One of the newly suggested roles of corpora was translator training. *"Corpora for translation studies have unique advantages over dictionaries and other references, since they are equipped with automatic search function capable and thus can be conveniently used to retrieve large amounts of linguistic data." (Hu, 2013:22)* This testifies once again, the relevancy of the "Web as Corpus" approach. Everything converges, directly or indirectly, on the same undergoing idea, and that is that the use of digital corpora, and the Web as a corpus has more benefits than what the common Web-user is aware of.

The use of the Web as a corpus allows the translator (or the student) to get information about the contexts in which a words or a phrase appears thanks to the search function, to perform an analysis of the ratio of equivalence between the source and the target texts, has advantages such as the digitalization of texts, visualization of data, diverse perspective in analysis, and the validity and reliability of research findings. (Li, 2007)

Parallel corpora were the type of corpora more commonly used in corpus-based translation studies. Malmkjær was the one to set the grounds for its use in

translation studies. She studied the advantages and limitations of the use of parallel corpora in translation studies. One of the most notorious disadvantage Malkmjær outlines was related to the use of parallel corpora in translation studies. She argued that the concordance lines generally used as an analytical tool do not always offer enough linguistic context for investigating features of entire texts, so that a certain aspect of the translation may be lost of blurred along the process. (Hu, 2013)

Besides parallel corpora, multilingual corpora were used to typify and describe the different translated texts, and this allowed to designate specific parameters that served to identify and compare different completely different translated texts, a thing that would not be possible before. "*Corpora can be used in the investigation of translator's style and (…) translation features [which typically occur in translated text rather than original utterances and are not the result of interference from specific linguistic systems]*" (Baker 1993:243) And as a result of combining a corpus-like methodology with translation studies, Baker was the first to present a list of features known as "features of translation" (Baker, 1993:243-273), which included the following:

- *– Explicitation, in the form of shifts in cohesion (Blum-Kulka, 1986) and insertion of additional information in the target text (Baker, 1992)*
- *– Disambiguation and simplification (Vanderauwera in Baker, 1993:243-247)*
- *– Textual conventionality in translated novels (Vanderauwera in Baker, 1993:243–247) and interpreting (Shlesinger in Baker 1993:243–247)*
- *– A tendency to avoid repetition present in the source text (Shlesinger in Baker, 1993 :243–247; Toury in Baker,1993:243–247)*
- *– A tendency to exaggerate features of the target language (Toury in Baker, 1993 :243–247; Vanderauwera in Baker, 1993:243–247)*
- *– Specific distribution of lexical items in translated texts vis-à-vis source texts and original texts in the target language (Shamaa in Baker, 1993:243–247)*

One could argue that these features would lose its valid empiric value once translators were consciously aware of them, as their translations would deliberately be conditioned by them. However, that is not completely true, as *"(…) the use of comparable corpora is also seen as a way of investigating aspects of translators'*

use of language which are not the result of deliberate, controlled processes. Translators may not be aware of these processes, but the translation product may provide indirect evidence of cognitive processing inherent to translation." (Olohan, 2013:423). For example, Olohan and Baker (2000) studied the use of the optional "that" together with reporting verbs "say" and "tell". The study concluded that the use of the optional "that" was "(…) *considerably higher in the Translational English Corpus than in a comparable corpus comprising texts from the British National Corpus*" (Olohan, 2013:423). And the results were correlated to one of the translation features listed above, explicitation, to be more precise (the first one). What explicitation means is that translators will usually prefer the use of longer forms rather than shorter constructions which may leave room for ambiguity in the translation. This is closely related to cognitive complexity, the addition, deletion or replacement of elements in the translation to make it as explicit as possible. Therefore, the higher the use of "that" in the translation, the more cognitive complexity and explicitation present in the translation task.

So, as it can be seen with this example, these "universal translation features" did not only give birth to a whole new era within translation studies, it also allowed its framework to expand and become more inclined to be linked with other academic fields, such as cognitive linguistics or psycholinguistics, to mention a couple.

Later, Baker decided to simplify these universals even further, classifying them into simplification, explicitation, normalization and leveling out (Baker, 1996) The universal of simplification is pretty self-explanatory, it refers to the translator's conscious, and unconscious efforts to simplify the information in the source text.

As aforementioned, this universal may be closely linked to psycholinguistics, cognitive linguistics, and pragmatics. Explicitation is the tendency of the translator and the translation to make explicit what the implicit meaning in the source material, or to increase the level of explicitness in the target by adding their own explanatory notes. This is, in turn, also related with simplification in a way. This universal is closely related to entailments, presuppositions, to pragmatics and semantics. The other two universals left are not as self-explanatory as these two. Normalization is the tendency of the translated text to conform or exaggerate the typical features of the target language (Baker, 1996:185).

And the last one, leveling out is "*the tendency of translated texts to gravitate around the center of any continuum rather than moving towards the fringes*" (Hu, 2013: 6).

3 Creation of a Corpus

The process of compilation and arrangement of the data for the corpus is very time-consuming and challenging for a single individual. In corpus linguistics there are not many studies with corpora created from a single researcher, the amount of workload would be too much for a single researcher, even more if he is a part-time researcher.

In formal corpus linguistics, it is assumed that the corpus presented has been compiled and analyzed by more than one person. A detailed study of a corpus is not something that can be done alone, it the field of corpus linguistics it is a given that for a corpus to be competent, it has to be done by a research team, there are too many aspects to be taken into an account. This is a crucial hindrance for users, such as freelance translators or students that want to compile a corpus by themselves. To come up with solutions to the problems of excessive length an effort to create a single corpus, methods of compilation are being steadily proposed, each new one cutting the time and effort required in the process of corpus creation and its subsequent analysis. This new methodology is closely linked to the Web as Corpus approach, as all the new advancements are done with the use of the Web. This type of methodology, that helps creating and compiling corpora using the contents in the Web, is done easily and at the moment, it is a "*right here, right now methodology*". Aston (1990) coined the methodology that uses the Web as the *ad-hoc* methodology. This was in the 90s, and since then a lot of things have happened in the Web. Aston's initial proposal has been overtaken by more contemporary concepts, such as virtual corpus (Ahmad, 1994), special purpose corpus (Pearson, 1998); *corpus especials* (Sánchez-Gijón, 2003); customized corpus (Austermülh, 2001); disposable [corpus] (Varantola, 2000); do-it-yourself (DIY) corpora (Zanettin, 2002). The user may use any of these phrases to refer to a corpus, but the common feature that all these different concepts have is an essential one: all of them need the Web. In a way, all of them are enclosed in the Web as corpus approach.

The first step in order to create a corpus is the compilation of data. Every corpus created using the Web starts with a simple query *"(…) searching a word can be compared to the first step in corpus creation.*" (Stubbs, 2007:7). This process may appear as the most tedious one for the average user when s/he has to create a corpus, but in reality, it is not that hard. The Web is an untapped source of data, accessible in a click of a mouse, for better or for worse as this may be. The average user will be able to find a considerable amount of information about the specific topic he is interested just by typing it in the search engine. Search engines, as previously indicated, are a double-edged sword. It is important to emphasize that search engines do not discriminate information, the Web is a huge anarchic database, and the only function of search engines is acting as a bottleneck. Search engines will not take into account the quality of the content, the format, or whether if it is a duplicate or not. Not all data that the search engines show as results after the initial query will be reliable or useful for the specific purpose of that corpus. In other words, *"While working with such a huge amount of data greatly enhances the chances of finding information about any item, this advantage is counterbalanced by the fact that, in many cases, dealing with too much data may result in the impossibility of assessing the real significance of the quantitative data retrieved*." (Gatto, 2013:76) The process of compilation can be very the most critical, in that if the data compiled in the first stages of the corpus is not reliable or valid, then the results will be skewed and so, they not be representative, therefore are the efforts of the user will be all for naught.

Another hindrance about using search engines is that they are not designed for linguistic purposes. "*The results (…) are displayed in a format which is not suitable for linguistic analysis, and results are ranked according to algorithms which escape the user's control")* (Gatto, 2013:75)

Zanettin is able to accurately encapsulate the dichotomous scenario presented so far can in this study with this statement:

> "*The first [problem] concerns procedures for assessing relevance and reliability: Information is dispersed in the Web through vast quantities of documents, and it is thus crucial for the translator to retrieve this information in the most efficient and effective way. The*

> *second relates to strategies and techniques for searching electronic texts: Search engines provide access points to Internet documents either through lists generated by full text searches or by pre-selected lists organized by topic and are thus catalogues rather than corpora.*" (Zanettin, 2002: 241)

This statement goes back to 2002, 17 years ago… Since then, this scenario has evolved at an exponential rate. The most common search engines nowadays are do not look like they were 10 years ago. In the past, search engines were designed using a query-based algorithm, in other words, the results were only exact matches of the query the user looked for, but nowadays their focus has shifted away.

The contemporary approach to search engines is more context-driven, the algorithm is designed to provide information estimated by the machine to be useful for the user (Broder, 2006). The ultimate goal for computational engineers is to completely revamp the figure of the search engine, to build a search engine as if it was an independent AI, an entity in and of itself. In other words, the new search engines "*do not simply 'passively' retrieve the information required but rather 'actively' supply (unsolicited and often commercial) information, not only in the form of 'banner ads' and 'sponsored links', but also through algorithms aimed at behavioral and contextual targeting*" (Levene 2010: 152).

But how do search engines actually search for the data that is to be compiled? Search engines usually have internal and exclusive Web crawlers which are not accessible to the average user, in other words, the user will not be able to exploit Google's Web crawler unless he uses Google. That is the marketing strategy of most search engines. The crawler, in a matter of milliseconds (See Figure X), comes back to the user with "About 1.240.00 results" to the user query, including verbatim and other contents that the Web crawler deems relevant for the user.

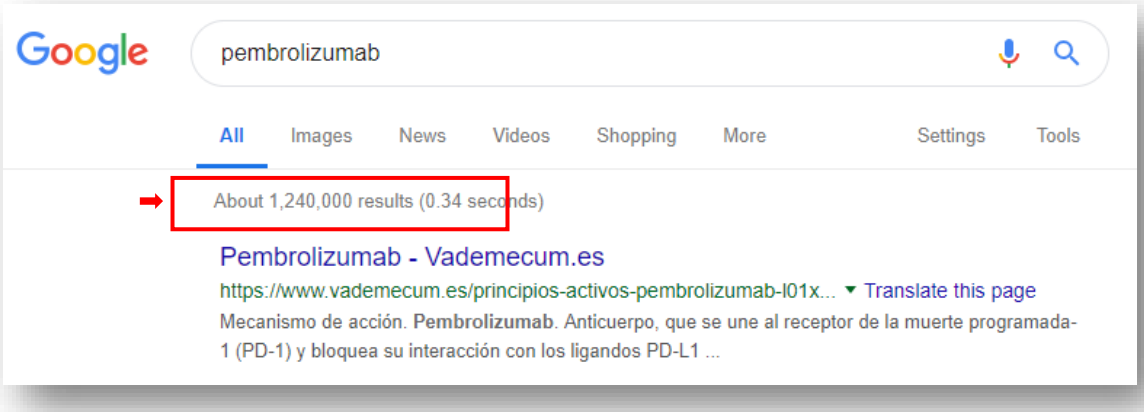


Figure 1. Google query interface

On account of the business model prevalent in most search engines, the linguistic aspects of the query are not the most relevant factor in the search logarithm of search engines. Generally, the factors that they take into account, at a very basic level, are "*the popularity of the page, measure (measured by the number of other pages linked to it), the number of times the search term occurs in the page, the relative proximity of search terms in the page, the location of search terms (for example, pages where the search terms occur in the title page get higher ranking), and even the geographical provenance of the query (which may prompt a bias to ranking higher those Websites which are closer to the user)*" (Brin and Page, 1998 in Gatto, 2013:77). The amount of processes going on behind the users' back is astounding. Google offers an advanced search, and it acts as a replacement for Boolean operators (AND, OR, NOT) that were used in former search engines. For instance, the user, instead of looking for *"intravenous AND oral*" in the search box (which come up with results not useful at all for the user, he will use the box "all these words" which represents the Boolean AND. The next option, "the exact word or phrase" is used to look exclusively for *verbatims* and words or phrases exactly in the order specified. The third option, "any of these words" represents the Boolean OR, and the fourth option is "none of these words", representing the Boolean operator NOT. As an illustration, if the user wanted to create a corpus focused on the medical/pharmaceutical field, as soon as the user looked for 'cancer' as a query in the search engine, the results would be inconclusive, to say the least. If the user

had to create a corpus based on that query, it would contain a mix of medical texts and horoscope related texts. Therefore, one possible solution would be using the Boolean NOT (-) to specify the field of the query by ruling other related expressions. For instance, “cancer – horoscope -zodiac” to cut out all the results related to the horoscope, or “cancer -disease -patients” the other way around.

In the case of Google, the user does not need to access this advanced search option every time the user decides to do a carefully detailed query, instead he can user simpler symbols such as + (for AND), “ “ (for ‘exact phrase match’), OR (for OR), - (for NOT). Another thing that might be useful for the user to know is that most of the contemporary search engines work with the Boolean AND by default, in other words, using a search engine you could look for a query in 3 different ways:

- antibody antigen
- antibody AND antigen
- antibody + antigen

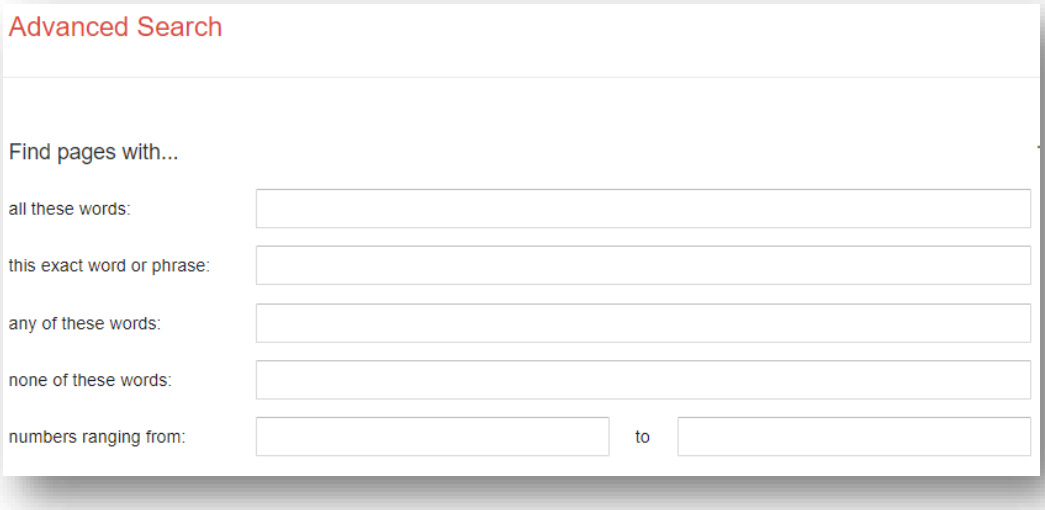


Figure 2. Google’s advanced search interface

Apart from Boolean specification, Google offers the user more tools to further narrow the query results, as it may be seen in Figure 3.

Figure 3. Additional Google's advanced search options

This way, the user can restrict the results to a single language, therefore keeping the noise in other languages in the corpus to a minimum. That way, if the user limits the results to English, the noise in German (as "die" is an article in German) will be minimized. This option, as basic as it may sound, is very handy for linguists, as this kind of noise is very common when the task of compilation is done automatically done by a search engine. Another option that may be handy for the user is the restriction of domain "site or domain". It may be not very evident at first, but limiting a search to a specific domain may increase the reliability of the corpus and make the results more significative. Mair (2007, 2012) and Cook & Hirst (2012) provided the necessary evidence to ascertain that corpora which used more national top-level domains, such as *.uk., .es, .fr,* or reserved domains for academic and divulgation content such as *.gov, .edu., .org* boost the reliability of the results. A corpus containing these types of domains is *"(…) more similar to a corpus known to contain texts from authors of the corresponding country than to a corpus known to contain documents by authors from another country"* (Cook and Hirst 2012: 281). To sum up everything that has been described in this last section, this toolkit that the search engines offer the users greatly streamlines the compilation task to the users, and not only that, but also the barriers and problems of the corpus related to "*representativeness and unreliability (…) are partly removed and the potential of the Web as a ready-made corpus is greatly enhanced."* (Gatto, 2013:87).

Language is constantly fluctuating, changing and evolving, and this is one of the major linguistic advantages this 'Web as Corpus' approach provides to professional translators, students and researchers. Thanks to the Web, and using some of the

methods mentioned below, it is possible to create an *ad hoc* corpus for a very specific linguistic target in a very streamlined method, which can be used for the task and hand and disposed shortly after. For example, a translator is asked work on a project about the state of the treatments for the non-invasive pancreatic cancer, for instance. That is not a topic that can be learned and understood just by reading one or two articles. The translator will need a tool to make his work simpler and smoother, as time is of the essence in this line of work, and that is one of the most noticeable applications of the Web as Corpus approach. The translator will be able to create in just a few steps an easy to use, and disposable corpus

4 How to Create a Corpus from the Web

In this section, different methods to build corpora using the Web will be explained step by step, in order to provide a practical view of all the theoretical aspects that have been developed up until now. The tools that have been selected are BootCaT*, Sketch Engine and iWeb Corpus*, as they are the three most accessible and renowned tools for this task at the moment. It is important to note that this section, and this work in general, solely intend to present how an *ad-hoc* corpus can be built using the Web. The posterior analysis of it, using a concordancer such as *Wordsmith Tools* or *AntConc*, or using the Web itself, although being very close and associated to the corpus created, will not be included in the posterior explanations.

Evaluating the performance of an unsupervised algorithm found in the web is hard. How can the user decide whether a page or a term used is pertinent for a corpus? How can the user review the quality of the list of Web pages obtained from the query? These are questions that have been considered in academic works dealing with the empirical side of the corpora field and could be further developed into a future study. The qualitative and quantitative study of the corpus build form the Web, although being correlated to the topic at hand, are not pertinent in this introductory study.

4.1 How to Build a Corpus Using Sketch Engine

In order to show how to create a corpus *ad-hoc* using Sketch Engine, one of the most functional and intuitive tools that can be found in the Web, the whole process of creating a corpus from the Web will be presented hereafter.

The first step that the user needs to select once s/he accesses Sketch Engine (through an academic account or the free trial version) is selecting the option “New corpus”, in order to create a new corpus from scratch. Sketch Engine offers corpora that are already made from a wide range of languages, but the nature of those corpora is too general and would not help too much a translator needing a specialized corpus.

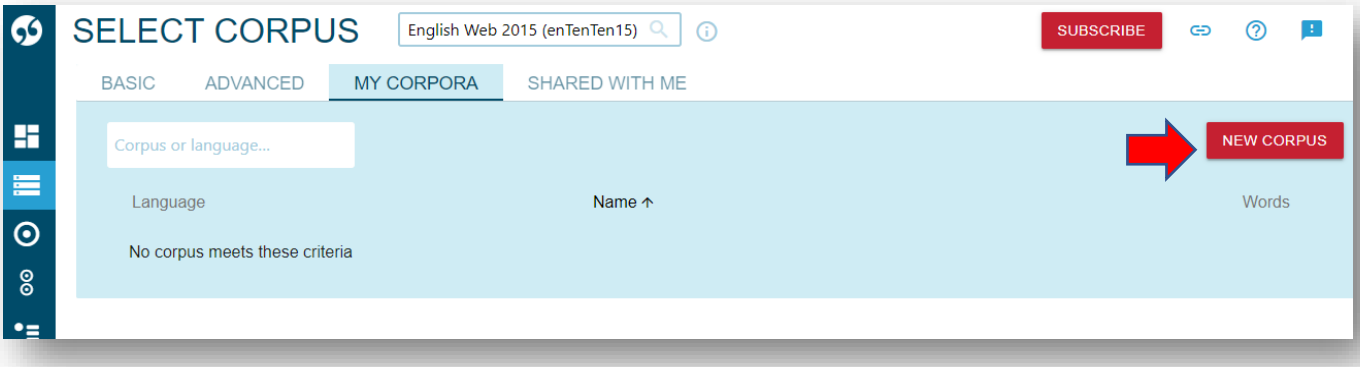


Figure 4. Initial interface to create a corpus in Sketch Engine

Once having selected this first option, the process is very streamlined and simplified

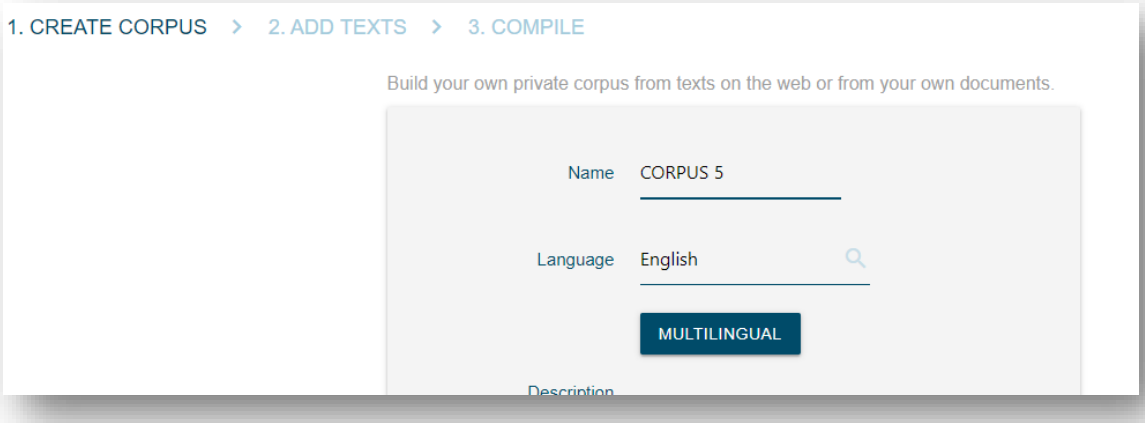


Figure 5. Name and specify the language of your corpus

You need to give your corpus a name, select the language of the corpus (in case it is monolingual) and it the user wished to do so, some description can be provided too.

And this is the step where the Web as Corpus approach comes into play. Sketch Engine allows the user to add data to their corpora by having the custom *Sketch Engine's* search engine find relevant texts on the Web for your corpus. If the user already has done their work and found texts about the topic of the corpus beforehand, they can also upload it to that corpus. The two methodologies can be combined into a single corpus.

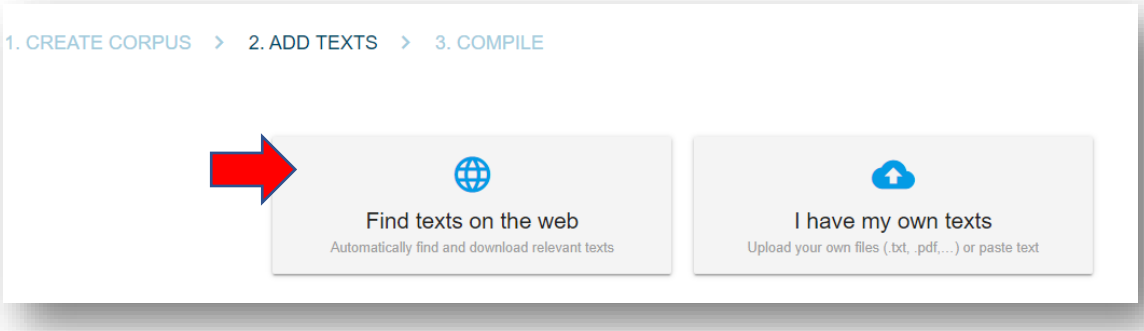


Figure 6. The two sources of texts available in Sketch Engine

Once the option "Find texts on the Web" is selected, the next screen that will appear will offer the user different methods to acquire the data from the Web. With the first option, "Web search", the user will have to input some words or phrases that define or represent the topic of the corpus (between 3 and 20 words). This query will be processed by the Bing motor engine, a direct competitor of Google, and the Web pages that Bing returns will be downloaded and processed into the corpus. This option is the most convenient one if the user is completely new to the topic or does not have time to research and look for some trustworthy sites beforehand. In "Web search" the workload is completely automated, and the user is not a part of the process of data mining at all. This option will exclusively download the content of the URL provided.

The remaining option has a very similar nature, it also demands the user to enter the URL to be downloaded, but in this case not only the URL will be downloaded, but also the complete Website in which the URL is included. In other

words, this option is used to download sections of the same Website. Consequently, this is the option that may present the most notable advantages and drawbacks. The drawback of this option is that as all the sections of a Web site are downloaded, there may be data which is not relevant at all for the task at hand, therefore filling the corpus with unnecessary data that may interfere and lengthen the work of the user.

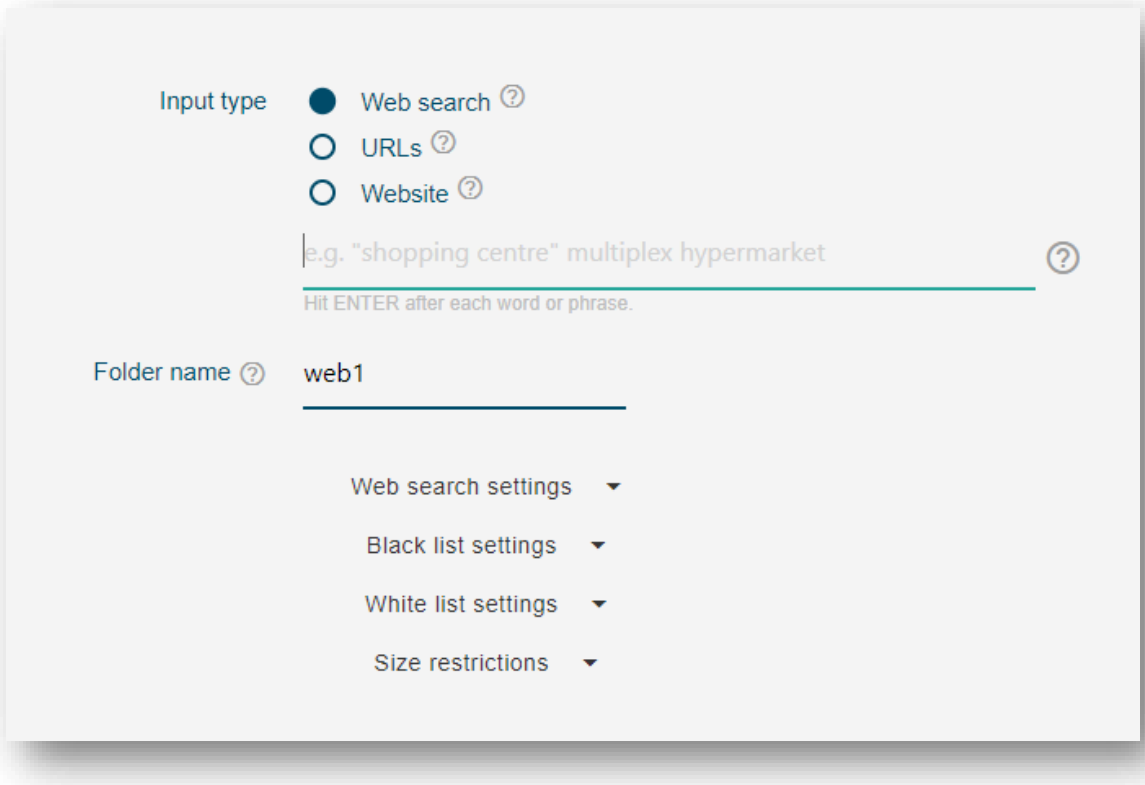


Figure 7. Methods of acquiring data from the Web.

With the Web Search option, Sketch Engine combines the input words into random groups of 3 and submits them to Bing. Then Bing searches the Web and sends the addresses of matching Web pages to Sketch Engine. Afterwards Sketch Engine downloads the pages and removes advertising, navigation menus and other linguistically content that may impede a swift download. This filter is used to keep the data compilation process as accurate as possible. More queries will produce more Bing searches, and therefore, a larger corpus, but the topic coverage and/or accuracy may be too wide. Fewer queries however, entail lesser Bing searches but generate a more compact and focus-oriented corpus.

Once the corpus is automatically compiled, the user can check the size of the corpus, the number of words and sentences in it and other linguistic information.

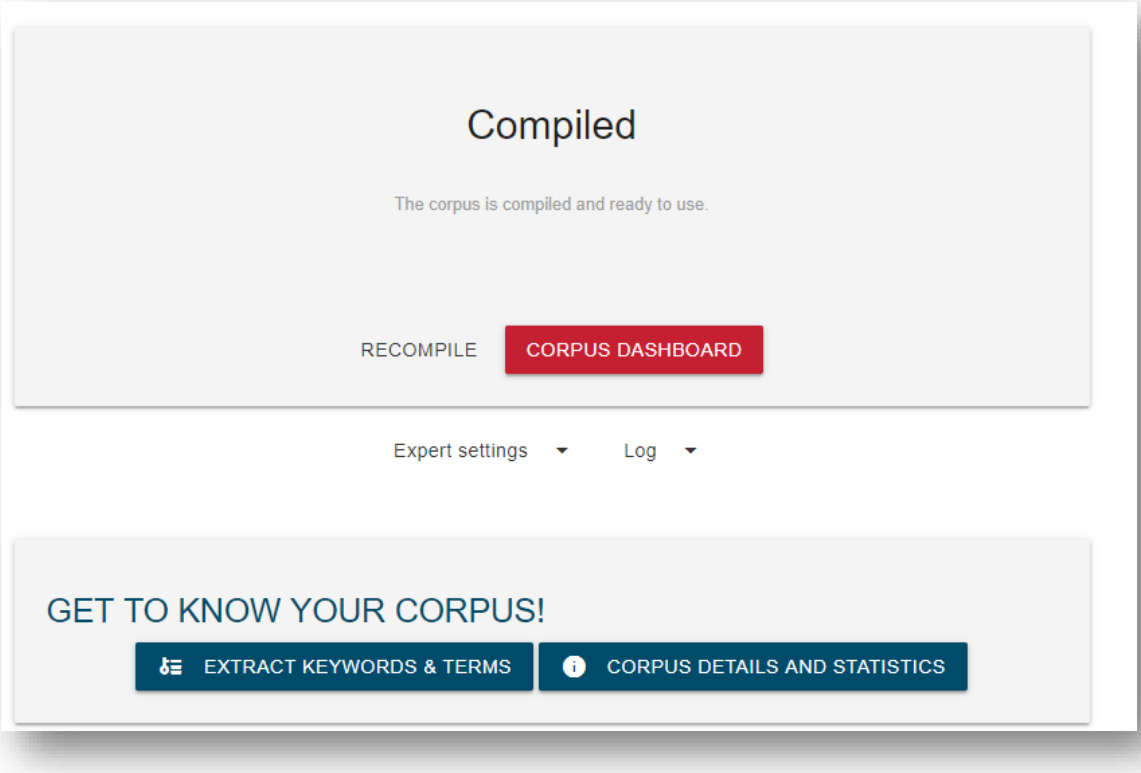


Figure 8. Your corpus is compiled

The 'Extract Keywords & Terms" will be used to check whether the texts in the corpus are really related to the intended topic of the corpus. This will allow the user to check if the data comprised in the corpus cover the expected topics.

Another appealing point to use Sketch Engine is the possibility to edit your corpus on the fly. The user can make it bigger by adding new texts, or smaller, if some of the texts included in it are do not fit.

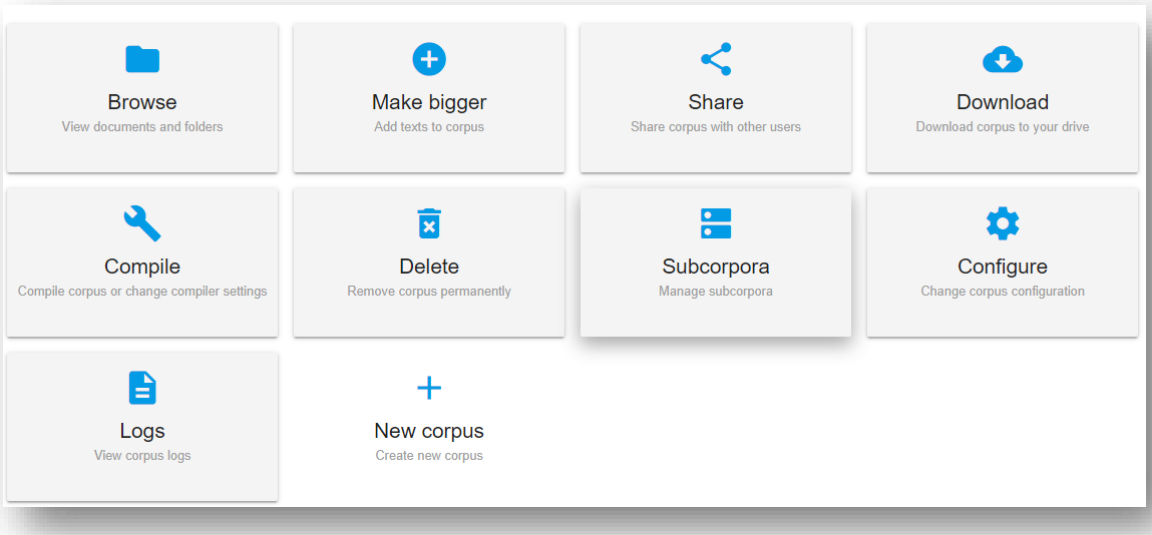


Figure 9. Final interface and functions for your corpus

4.2 How to Build a Corpus Using BootCaT

BootCaT is a kit of programs designed to create specialized corpora using the Web. This toolkit was not initially intended to be made available to the public use, so its original interface and accessibility have been greatly simplified to become more user-friendly. In a similar fashion to the previously described *Sketch Engine*, or to *iWeb Corpus, BootCaT* relieves the user from looking for the queries individually, downloading the results and the tedious format changes. It is a toolkit designed to "tool to help language professionals build the corpus they need, whenever
they need it and as quickly as possible." (Gatto, 2013:140)

The only thing that the user will need to do to initiate the process of creating a corpus is selecting a number of key terms relevant for the topic of the corpus. These key terms receive the name of "seeds", and even though they are only the first step towards getting a corpus, they are very significant of the final result of the corpus

> *"When compiling a specialized corpus from a text database by use of a query, there is a trade-off between precision and recall (e.g. Chowdhury 2004: 170). That is, there is a tension between,*

on the one hand, creating a corpus in which all the texts are relevant, but which does not contain all relevant texts available in the database, and, on the other, creating a corpus which does contain all available relevant texts, albeit at the expense of irrelevant texts also being included. (Gabrielatos, 2007:6)"

Based on the seeds that the user has chosen, the system automatically downloads the top results of the selected search engine. Contrarily to *Sketch Engine, BootCaT* allows the user to choose the search engine he wishes to use, and it is not restricted only to Bing.

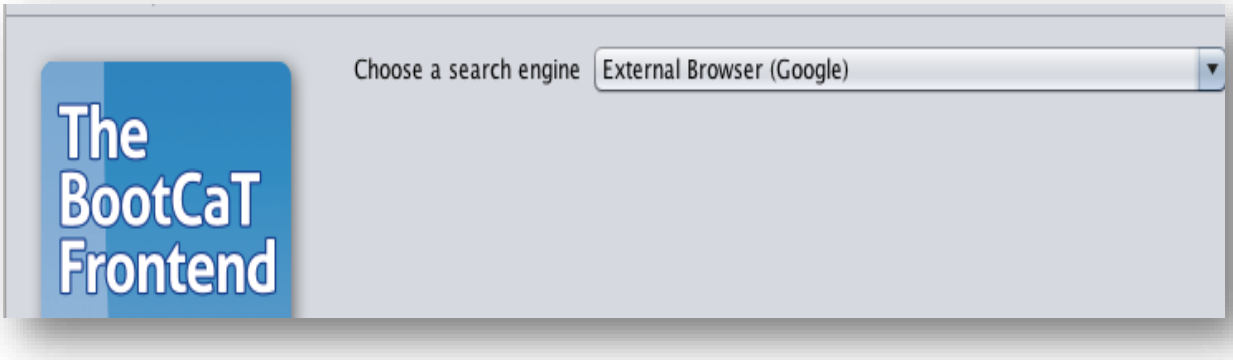


Figure 10. Specify the search engine for the data compilation

Once the search engine has been selected, the user will provide the seeds that will be used to generate the queries that will be submitted to the search engine. The minimum number of seeds that the user needs to provide is 5. If we had to compare this same step to Sketch Engine, we could say that in terms of the interface, this one is a little bit more intuitive.

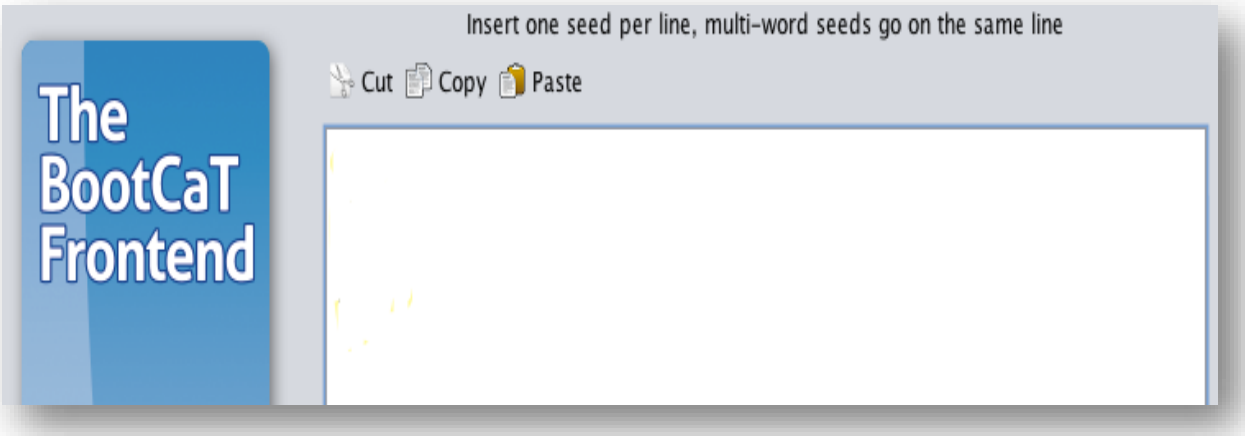


Figure 11. Insert the seeds for the search engine

These seeds will be combined into "tuples", which basically mean combination of words. With BootCaT, the length of the tuples can be modified, while in Sketch Engine this was not possible. BootCaT has a specific step focused on the generation of tuples, whereas in Sketch Engine this was done automatically and, in the background, or at least it was not as accessible as what BootCaT offers. There is an actual step to generate the queries, they are not done fully automatically.

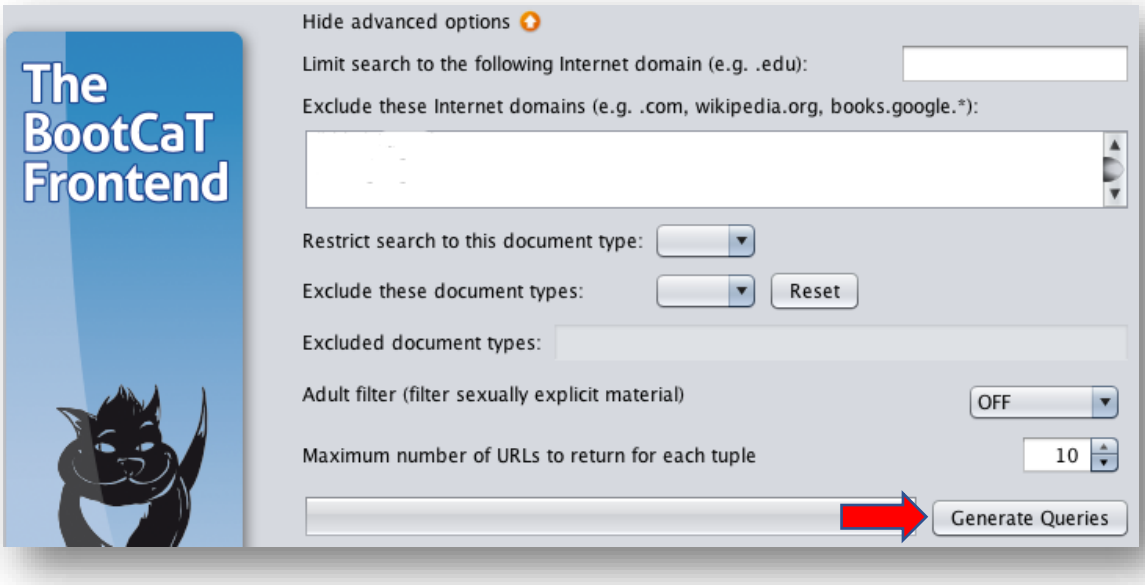


Figure 12. The user needs to select "Generate Queries".

The user needs to actively select the option “Generate Queries” after he has tweaked all the previous options. Once the queries are generated, BootCaT provides a list of potentially relevant URLs that are the results of the queries. At this point the user has the option of inspecting the URLs and trimming them; the actual Web pages are then retrieved, converted to plain text and saved in "txt" format.

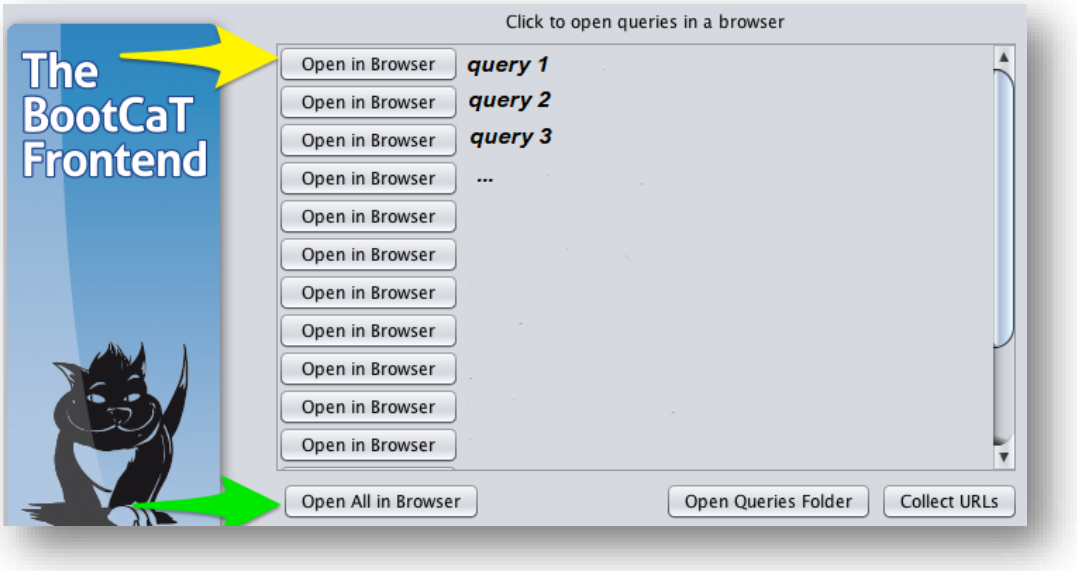


Figure 13. List of queries.

A distinctive feature of BootCaT is that it allows the user to open all the queries that it has made in the browser, so the user is able to identify suspicious URLs or directly see if there is any URL that may not be in line with others. Once the user is ready to proceed to the compilation of the URLs, the only thing to do is select the option “Collect URLs”.

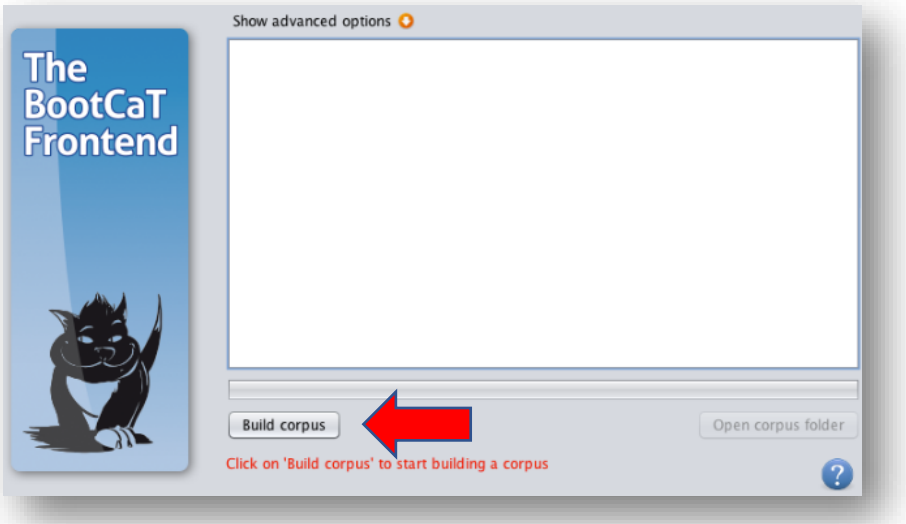


Figure 14. Next, the user needs to select "Build Corpus"

Finally, when the user selects "Build corpus", not only the URLs will be downloaded, they will also be cleaned of menus, navigation bars, ads, disclaimers and automatic error messages that may compromise the corpus. It is important to take into account

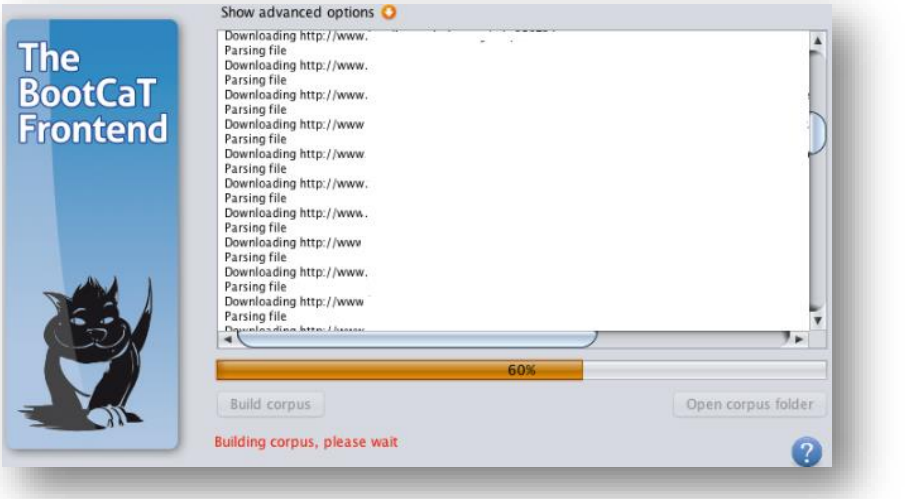


Figure 15. The corpus is being created.

that this process is completely automated, so the filter is not the definitive solution against unwanted data. Some unwanted elements may still be present in the corpus.

Once the download is complete, a new window will appear displaying the contents of the folder where the corpus data is stored.

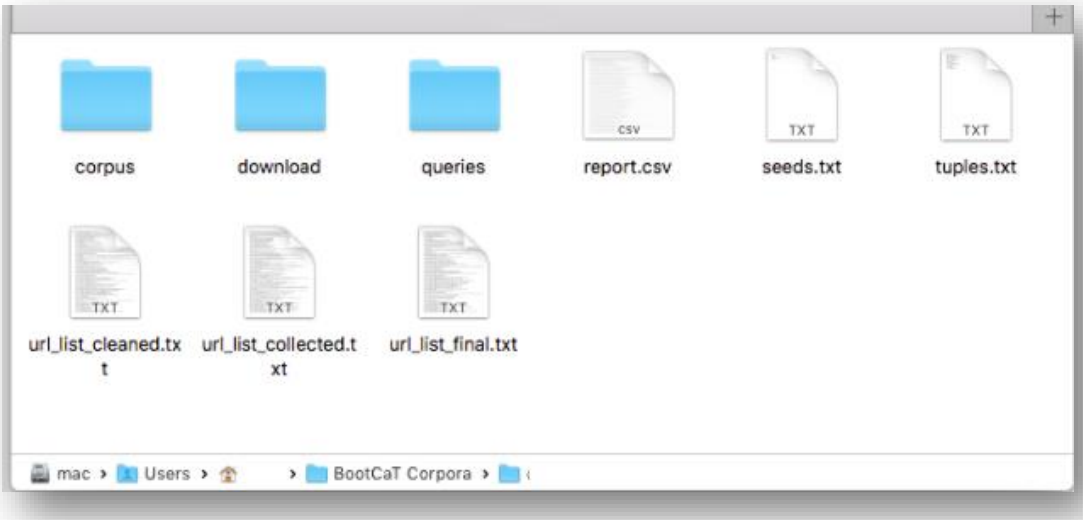


Figure 16. This is how the folder containing the corpus will look.

4.3 How to Build a Corpus Using the iWeb Corpus

The *iWeb Corpus* is one of the most recent of corpora released on the Internet (May 2018), and despite its recent publication, it is the biggest and most flexible corpus available on the Web. When it comes to its size, it has 14 billion words, in other words, it is 25 times bigger than the Corpus of Contemporary American English (COCA) and 100 times bigger than the British National Corpus (BNC). See Figure X for a visual illustration of its exorbitant size.

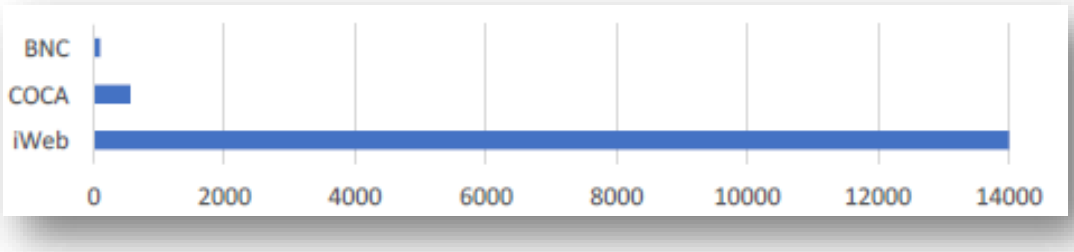


Figure 17. Comparison of size between corpora (in words)

The argument of its sheer size is a pretty appealing one by itself, and we cannot forget that it grows by the day, making *iWeb* a gigantic public corpus, but that is not the only thing that this novel corpus brings to the scene.

Another one of the features that make the iWeb such a valuable asset for the iWeb Corpus is that it lets the user create an online corpus in just 4-5 seconds for any topic that comes into mind within the iWeb Corpus. In other words, the user is making a sub-corpus inside the iWeb Corpus. Everything is automated, it is extremely simple to do so, and the only thing the user has to do is search for a keyword for the topic of the future corpus. For instance, let's say that the user wants to make a corpus about "*cancer*". The first step is using the iWeb Corpus as if it was a plain search engine such as Google. (See Figure 18)

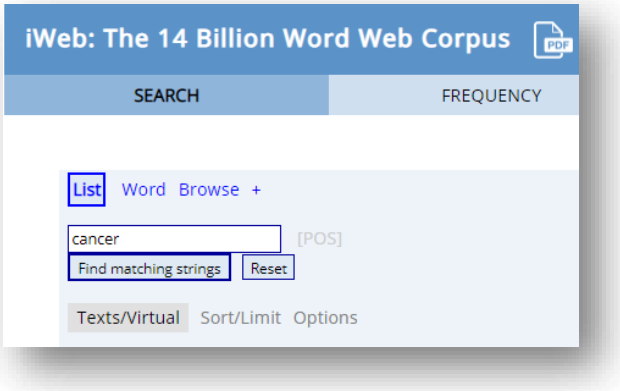
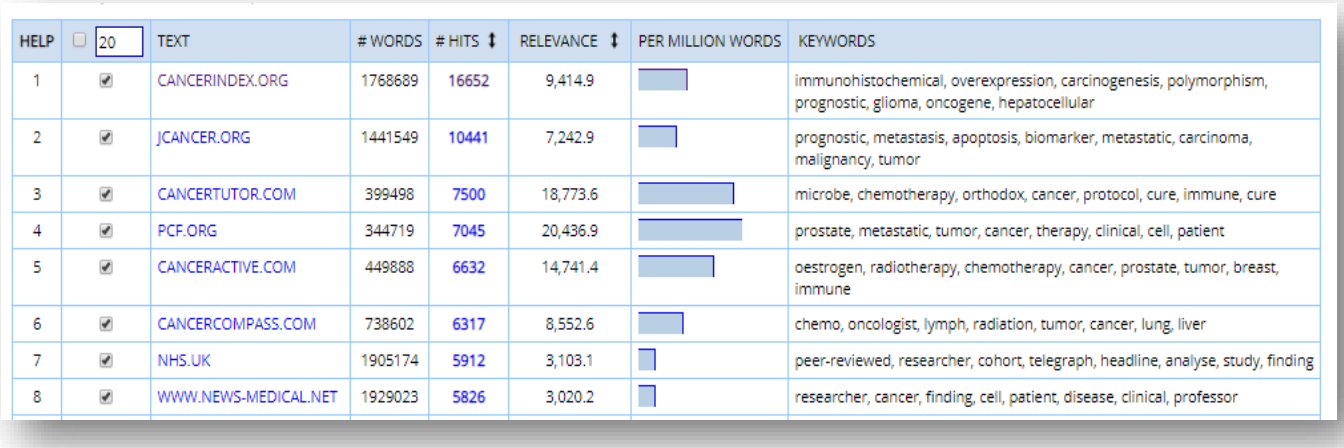


Figure 18. The first screen the user will find in the process of compiling a corpus

Once the query is done, the next screen presents the user a screen similar to Figure 19.



HELP	20	TEXT	# WORDS	# HITS	RELEVANCE	PER MILLION WORDS	KEYWORDS
1	☑	CANCERINDEX.ORG	1768689	16652	9,414.9		immunohistochemical, overexpression, carcinogenesis, polymorphism, prognostic, glioma, oncogene, hepatocellular
2	☑	JCANCER.ORG	1441549	10441	7,242.9		prognostic, metastasis, apoptosis, biomarker, metastatic, carcinoma, malignancy, tumor
3	☑	CANCERTUTOR.COM	399498	7500	18,773.6		microbe, chemotherapy, orthodox, cancer, protocol, cure, immune, cure
4	☑	PCF.ORG	344719	7045	20,436.9		prostate, metastatic, tumor, cancer, therapy, clinical, cell, patient
5	☑	CANCERACTIVE.COM	449888	6632	14,741.4		oestrogen, radiotherapy, chemotherapy, cancer, prostate, tumor, breast, immune
6	☑	CANCERCOMPASS.COM	738602	6317	8,552.6		chemo, oncologist, lymph, radiation, tumor, cancer, lung, liver
7	☑	NHS.UK	1905174	5912	3,103.1		peer-reviewed, researcher, cohort, telegraph, headline, analyse, study, finding
8	☑	WWW.NEWS-MEDICAL.NET	1929023	5826	3,020.2		researcher, cancer, finding, cell, patient, disease, clinical, professor

Figure 19. The main results of the query will look like this

This screen shows all the different sources that iWeb Corpus has found in its database, the user can select the websites that interest him the most. If s/he desires to do a corpus only with Websites with an .org domain, then s/he is able to select and deselect sources as he wishes using the tick box right next to the numbers. The sources are ranked by order of relevance. The more times the query appears in it, the higher it will be in the list. Once the user has selected all the desired sources he wants to include in his corpus, it is time to search within it as if it were its own stand-alone corpus. To search within the corpus that has been just created, the user only has to select the “Word” option and select the corpus that has just been created. Once this option has been selected, all the user’s subsequent queries will be done to the new corpus.

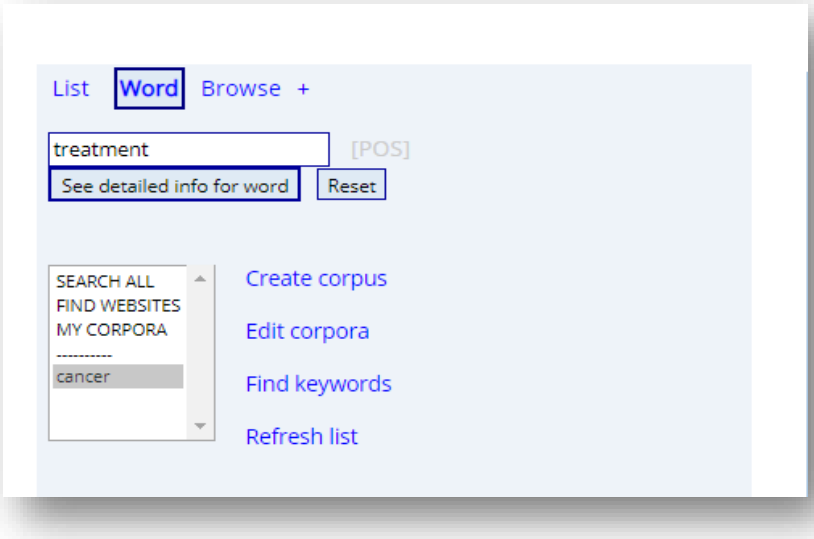


Figure 20. Looking for the word 'treatment' within the 'Cancer' corpus.

Each sub-corpus created by the iWeb Corpus has more detailed pages, such as collocates information, as seen in Figure 21, clusters, and a dictionary. The collocates page, for example includes information about which words appear together more frequently the query 'treatment'. As it can be seen in Figure X4, the Noun that appears together with 'treatment' the most is 'patient', the Verb is 'receive' and the Adjective is 'medical'

+ NOUN		NEW WORD	?
67798	4.01	patient	
67468	4.80	cancer	
58935	2.89	option	
42313	6.17	diagnosis	
37180	3.71	disease	

+ ADJ		NEW WORD	?
68695	4.16	medical	
43547	3.91	effective	
15320	2.86	appropriate	
13142	5.37	surgical	
11729	3.07	mental	

+ VERB		NEW WORD	?
52031	2.59	receive	
20121	2.80	seek	
15197	4.94	undergo	
6008	4.02	prescribe	
5104	3.33	administer	

Figure 21. Collocates information of the word 'treatment' in the 'Cancer corpus'

The iWeb corpus does not only offer linguistic empirical information such as frequency order, clusters, collocations and such, it goes one step beyond and uses all the tools that the Web has to offer. As far as information goes, the iWeb corpus is one of the precursors that is attempting to take the online corpora to the next stage. This online corpus makes the most of its virtual environment and harnesses

tools and functionalities that are exclusively available in the Web. One of these functionalities that the iWeb corpus offers is the dictionary page. This corpus includes a definition of the selected word or idiom to offer the user a better understanding of its meaning. Besides providing definition and the different forms of the words, iWeb also offers links to Google Images containing that word, or directly related to it, and it also includes the pronunciation of the word (PlayPhrase, YouGlish and Yarn). It also offers translational functionalities, as the user is able to translate the word to almost any language using four different sites (*Google, WordRef, Reverso and Linguee*). But that is not all. It also includes synonyms of the words, words with more specific meanings (hyponyms) and more general meanings (hypernyms) (i.e. plant, flower, rose). It is safe to say that the iWeb Corpus is the most complete tool when it comes to virtual corpora. It is not only the biggest corpus that can be found in the Web, it is also the corpus that offers the most tools to be used together with the corpus.

These functionalities make the iWeb the ideal corpus for researchers, teacher, and students alike. Learners will be able to see and listen how a word is pronounced and what does it actually mean (most regular corpora do not include definition, so this is a turning point for corpora). Teacher, similarly to students, will be able to approach corpora in a more practical way. And researchers will have data that will open new directions for research, like interacting with translations, synonyms, hyponyms or hypernyms.

5 Final Remarks

All of the tools that have been described previously, added to all the continuous technological breakthroughs plus the continuous academic developments help language professionals, students and researchers build the corpus they need, whenever they need it and as quickly as possible. All of the drawback that the 'Web as Corpus' approach will eventually be dealt with in the future, so the drawbacks that are clearly outweighed by the smoothness and simplification of the process of compilation of the data. Further works could be about the evaluation of the performance of the corpora created with the Web, and how these evaluations differ

depending on the tool used, or on how these drawbacks can be effectively avoided.

By making the creation of ad hoc temporary corpora an easily and achievable goal, all these tools really bring the ideal notion of the Web as a sort of virtual multilingual multipurpose corpus on demand a bit closer to reality. And all the new steps and features that are being added by the day (some already seen in the iWeb Corpus) open interesting new prospects for corpora linguistics. The 'Web as Corpus' is just another step in the evolution chain. The next step for virtual corpora is the synthesis of corpora with multimedia, such as videos or audios, to create a completely new type of adaptive corpora that would combine the features of traditional corpora with the new features that are exclusive of the Web and would bring corpora to a new level. The limitations for these types of corpora are still to be seen, as well as their potential. Corpora are the backbone of a lot of translations and machine translation projects, so it is an influential topic that most people are unaware of. The ultimate goal of this study has been bringing people closer to virtual corpora, and shed some light into them, and now more than ever, as the next stage for corpora is knocking at our doors.

6 Bibliography

- Aarts, J. (1999). '*The description of language use'*. H. Hasselgård & S. Oksefjell (eds) (1999). Out of Corpora: Studies in honour of Stig Johansson. Amsterdam and Atlanta:

- Baker, M (1993). '*Corpus Linguistics and Translation Studies. Implications and Applications'*. Baker, Text and Technology: In honour of John Sinclair. Amsterdam/Philadelphia

- Baroni, M. and Ueyama, M. (2006), '*Building general- and special-purpose corpora by Web crawling*', in Proceedings of the NIJL International Symposyum, Language Corpora: Their Compilation and Application, 31–40.

- Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009), '*The WaCky wide Web: a collection of very large linguistically processed Web-crawled corpora'*, Language Resources & Evaluation, 43, 209–26.

- Biber, D., Conrad, S. and Reppen, R. (1998), '*Corpus Linguistics. Investigating Language Structures and Use'*. Cambridge: Cambridge University Press.

- Broder, A. 2006), *'From query-based Information Retrieval to context driven Information Supply*'. Workshop on The Future of Web Search, Barcelona

-Cook, P. and Hirst, G. (2012), '*Do Web Corpora from Top-Level Domains Represent National Varieties of English?*', in *Proceedings, 11th International Conference on Statistical Analysis of Textual Data (2012*
Available at http://www.cs.toronto.edu/~pcook/CookHirst2012.pdf

- Crystal, D. (2006), '*Language and the Internet'*. Bloomsbury: Bloomsbury Academic
- Davies, M (2017). '*The new 4.3 billion word NOW corpus, with 4--5 million words of data added every day'* Available at: https://www.english-corpora.org/iWeb/help/iWeb_overview.pdf

- Ferraresi, A (2013) *Google and Beyond: Web-As-Corpus Methodologies for Translators*, Revista Tradumàtica, 7, Available at: *http://www.fti.uab.cat/tradumatica/revista/num7/articles/04/04art.htm*

- Fletcher, W (2004), '*Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora*', in G. Aston et al. (eds), 271–300.

- Francis, N. W. (1992), '*Language corpora B.C*.', in J. Svartvik (ed.), 17–33.

-Gatto, M. (2013). '*The Web as Corpus: Theory and Practice'* Cambridge: Cambridge University Press

- Grefenstette, G. and Nioche J. (2000), '*Estimation of English and non-English language use on the WWW'*. In Proc. RIAO (Recherche d'Informations Assisté par Ordinateur), 237–246, Paris.

- Hu, K. (2016) '*Introducing corpus-based translation studies'*. Springer.

- Hundt M., Nesselhauf, N. and Biewer, C. (eds) (2007), '*Corpus Linguistics and the Web'*. Amsterdam: Rodopi.

-Kennedy, G. (1998), '*An Introduction to Corpus Linguistics'*. London: Longman

-Kilgarriff, A. (2001), '*Web as corpus'*, in Proceedings of the Corpus Linguistics Conference (CL 2001), University Centre for Computer Research on Language Technical Paper, Vol. 13, Special Issue, Lancaster University,

- Kilgarriff, A. and Grefenstette, G. (2003), '*Introduction to the Special Issue on the Web as Corpus'*, in Computational Linguistics, 29, 3, 333–47.
- Laviosa, S.(2002), Corpus-Based Translation Studies: Theory, Findings, Applications. Amsterdam: Rodopi.

- Lawrence, S. and Giles, C. L. (1999), '*Accessibility of Information on the Web'* in Nature.

-Levene, M. (2010), '*An introduction to search engines and Web navigation*' New York: John Wiley & Sons.

- McEnery, A. M., & Wilson, A. (2001*). Corpus linguistics: An Introduction.* Edinburgh: Edinburgh University Press

-Oakes, M. & T. McEnery (2000). 'Bilingual text alignment – an overview.' S. P. Botley, A. M. McEnery & A. Wilson (eds) (2000). Multilingual Corpora in Teaching and Research. Amsterdam and Atlanta: Rodopi,

-Olohan, M. (2004) '*Introducing Corpora in Translation Studies'*. London/New York: Routledge.

- O'Keeffe, A. & McCarthy, M. (2010) *'Historical perspective: What are corpora and how have they evolved*?', The Routledge Handbook of Corpus Linguistics. London: Routledge, 3-13.

- Schafer, R. and Bildhauer, F. (2013). '*Web Corpus Construction'*. San Francisco: Morgan & Claypool.

- Sharoff, S. (2006) '*Creating general-purpose corpora using automated search engine queries*', in WaCky! Working papers on the Web as Corpus. Bologna

- Sinclair, J. 2005), '*Corpus and Text. Basic Principles'*, in Wynne, M. (ed.), Developing Linguistic Corpora: a Guide to Good Practice, Oxford: Oxbow Books, 1–16. http://ahds.ac.uk/linguistic-corpora

- Storrer, A., and Beißwenger, M. (2008) '*Corpora of computer-mediated communication*', in Corpus linguistics: An international handbook, A. L¨udeling and M. Kyto(eds.). Berlin: Mouton de Gruyter.

- Stubbs, M. (2007), '*On texts, corpora and models of language'*, in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds)

- Stubbs, M. (2004) '*Language Corpora*'. In Davies, Alan/Elder, Catherine (eds.), Handbook of Applied Linguistics. Oxford: Blackwell.

- Toury, G. (1995). '*Descriptive Translation Studies and beyond'*. Amsterdam and Philadelphia: John Benjamins.

- Ueyama, M. (2006) '*Creation of general-purpose Japanese Web corpora with different search engine query strategies'*, in WaCky! Working papers on the Web as Corpus. Bologna

-Zanettin, F. (2002), *'DIY Corpora: the WWW and the Translator'*, in B. Maia, J. Haller and M. Ulrych (eds), Training the Language Services Provider for the New Millennium. Porto: Facultade de Letras, Universidade do Porto
Available at http://www.federicozanettin.net/DIYcorpora.htm